MediX-R1: Open Ended Medical Reinforcement Learning

Anonymous ACL submission

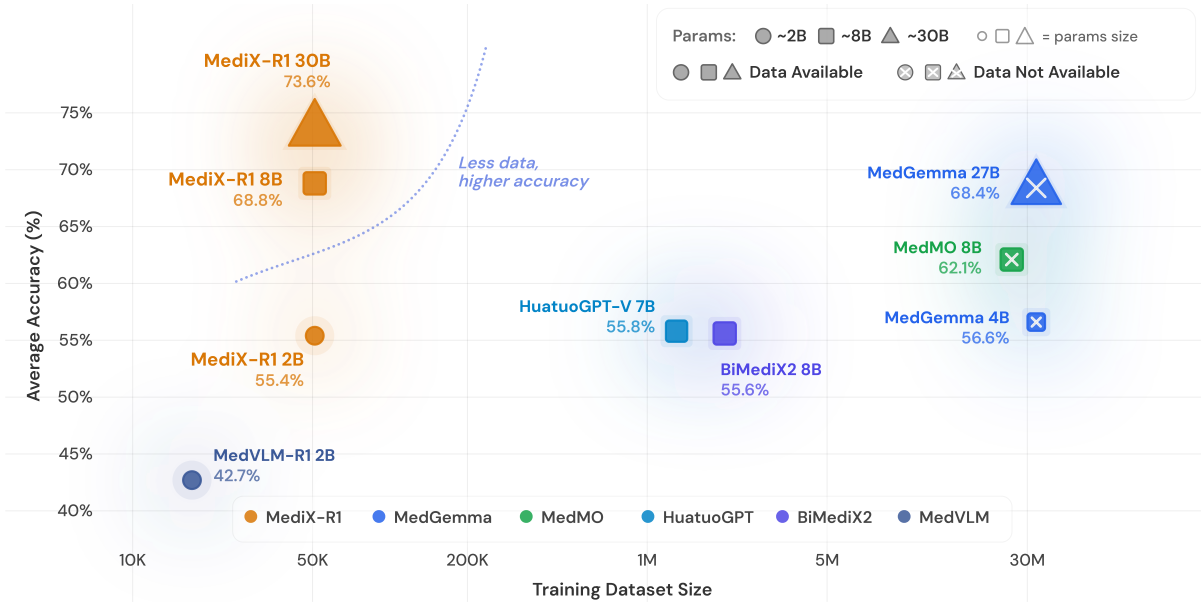

Figure 1: **Average accuracy across multimodal medical benchmarks vs. training dataset size for recent medical VLMs.** Colors denote model families; marker shape/size indicates parameter scale ∼(2B, 8B, 30B). × indicates the absence of public training data. MediX-R1 8B (68.8%) surpasses MedGemma 27B (68.4%) while using significantly less training data, and MediX-R1 30B achieves the highest overall accuracy (73.6%).

Abstract

In this work, we introduce MediX-R1, an open-ended Reinforcement Learning (RL) framework for medical vision language models (VLMs) that enables free-form answers beyond multiple-choice formats. MediX-R1 fine-tunes a baseline vision-language backbone with Group Based RL and a composite reward tailored for medical reasoning: an LLM-based accuracy reward that judges semantic correctness with a strict YES/NO decision, a medical embedding-based semantic reward to capture paraphrases and terminology variants, and lightweight format and modality rewards that enforce interpretable reasoning and modality recognition. This multi-signal design provides stable, informative feedback for open-ended outputs where traditional verifiable or MCQ-only rewards fall short. To evaluate performance, we employ a unified evaluation framework for both text-only and multimodal (image+text) tasks that uses a Reference-based LLM-as-judge in place of brittle string-overlap metrics, capturing semantic correctness, reasoning, and contextual alignment. Our results show that open-ended RL with comprehensive reward signals can effectively strengthen medical reasoning in multimodal models. Despite using only $\sim$ 51K instruction examples, MediX-R1 achieves strong performance across standard medical LLM (text-only) and VLM (image + text) benchmarks, outperforming existing open-weight medical VLMs, with large gains particularly on open-ended clinical tasks. All resources, including dataset and *training* and *evaluation* code, will be publicly released; *anonymized code* is available at **MediX-R1**.

1 Introduction

Large medical language and vision-language models are increasingly deployed for clinical question

Model	Diverse Medical Modalities	Single-Stage RL	Interpretable Reasoning	Open-Ended Responses	Annotation-Free Reasoning	Composite RL Reward
MedVLM-R1 *(MICCAI 2025)*	✗	✓	✓	✗	✓	✗
BiMediX2 *(EMNLP 2025)*	✓	✗	✗	✓	✗	✗
HuatuoGPT-V *(EMNLP 2024)*	✓	✗	✗	✓	✗	✗
MedGemma *(arXiv 2026)*	✓	✗	✓	✓	✗	✗
MedMO *(CVPR 2026)*	✓	✗	✓	✓	✗	✗
MediX-R1	✓	✓	✓	✓	✓	✓

Figure 2: **Differences with existing Medical VLMs:** MediX-R1 integrates diverse modalities, interpretable reasoning, and composite RL rewards, enabling practical clinical use.

answering, triage support, report drafting, and education (Chen et al., 2024a; Sellergren et al., 2025; Pieri et al., 2024). Many of these tasks are inherently open-ended: clinicians expect concise but free-form answers that can flexibly incorporate context, uncertainty, and multimodal evidence. However, most training and evaluation pipelines remain tailored to Multiple Choice Questions (MCQ) or string-matching regimes, which (i) under-reward valid clinical paraphrases, (ii) fail to measure reasoning quality or modality recognition, and (iii) do not provide reliable signals for reinforcement learning (RL) in open-ended settings. As a result, models trained only with supervised objectives or MCQ-style rewards often struggle to produce faithful, interpretable, and robust clinical responses across diverse medical tasks.

RL has improved reasoning in domains with verifiable rewards (e.g., math and code) as shown by DeepSeek models (Shao et al., 2024; Guo et al., 2025), but medical tasks rarely admit executable checks. Binary exact match is too brittle for clinical phrasing; BLEU/ROUGE can misjudge semantically correct answers; and free-form VLM outputs complicate visual inference. Moreover, using a single reward signal can induce instability or reward hacking, especially when the signal is noisy or overly permissive. Hence, it is desirable to have a systematic approach for training medical VLMs with open-ended RL that integrates semantic correctness with structural and modality constraints, while remaining data- and compute-efficient.

In this work, we propose MediX-R1, an open-ended medical RL framework that fine-tunes a baseline multimodal backbone using a group-based RL algorithm with a composite reward tailored for clinical reasoning and applicable across different RL optimization strategies (e.g., GRPO, GSPO, and DAPO). The proposed composite reward combines: (1) an LLM-based accuracy reward that enforces a strict YES/NO decision on semantic correctness, (2) a medical embedding-based semantic reward that captures paraphrases and terminology variants, (3) a lightweight format reward that elicits interpretable reasoning traces, and (4) a modality recognition reward that discourages cross-modality hallucinations by requiring explicit modality tags. This multi-signal objective stabilizes optimization and gives informative feedback where traditional verifiable or MCQ-only rewards fall short, enabling single-stage, open-ended RL on clinical tasks.

Differences with existing Medical VLMs: Figure 2 contrasts MediX-R1 with strong open models across key clinical capabilities. First, on *Diverse Medical Modalities*, MediX-R1 supports diverse medical modalities including X-Ray, CT, MRI, Microscopy/Histopathology, Ultrasound, Fluoroscopy, Endoscopy, Angiography, Mammography, Clinical Photography, SPECT (Single Photon Emission Computed Tomography), OCT (Optical Coherence Tomography), and Fundus imaging, whereas MedVLM-R1 (Pan et al., 2025) is limited to radiology images. Models like MedGemma (Sellergren et al., 2026), HuatuoGPT-Vision (Chen et al., 2024b), MedMO (Deria et al., 2026), and BiMediX2 (Mullappilly et al., 2025) provide coverage on clinical modalities but they require extensive multi-stage training. On *Single-Stage RL*, most baselines rely on multi-stage pipelines (pretraining $\rightarrow$ SFT $\rightarrow$ RL), whereas MediX-R1 is trained end-to-end with a single RL stage using our composite reward (Sec. 2.2). This simplifies training and, importantly, enables *open-ended* RL directly (unlike MedVLM-R1), because the Reference-based LLM-as-judge accuracy signal and medical embeddings provide reliable feedback beyond MCQ exact match. The composite design (format + LLM judge + embeddings + modality recognition) stabilizes optimization and reduces reward hacking (Fig. 5), translating into the best average performance in Table 1. For *Interpretable Reasoning*, MediX-R1 generates explicit reasoning traces en-

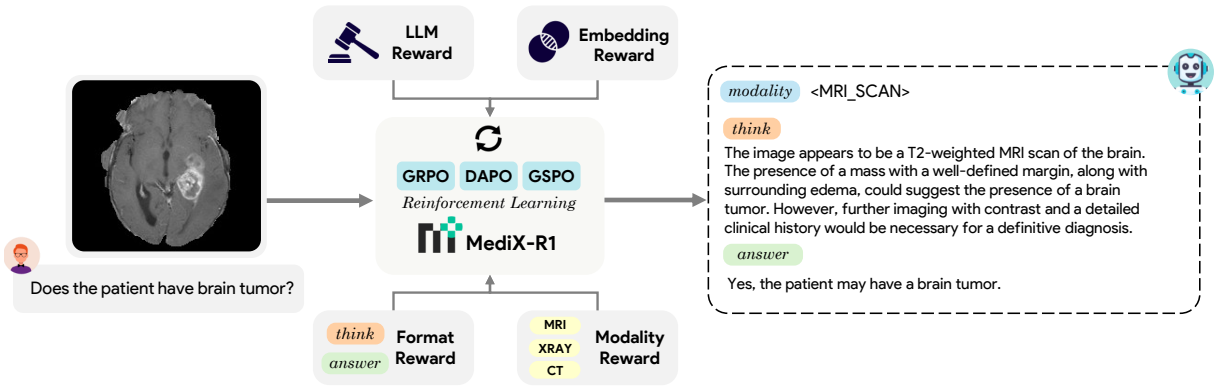

Figure 3: **MediX-R1: Overall Architecture** The MediX-R1 reinforcement learning framework for open-ended medical reasoning. An input of a medical image and a natural language question is processed by MediX-R1. The model's policy is trained using Group Based RL, which leverages a multi-faceted reward signal. This reward is composed of: a) an LLM-based reward for evaluating the quality and accuracy of the output; b) an embedding-based reward to ensure semantic alignment; c) a format reward to enforce the desired output structure (`<think>` and `<answer>` blocks); and d) a modality reward to ensure the response is grounded in the specified imaging modality. This reward-guided approach encourages the model to generate accurate and interpretable reasoning paths.

closed in `<think>...</think>`, enforced by the format reward, making the decision path auditable. Several baselines do not reliably produce structured clinical rationales. While multiple models support *Open-Ended Responses*, MediX-R1 is explicitly optimized for free-form clinical answering with modality recognition, which reduces cross-modality hallucinations and improves VLM robustness. MediX-R1 achieves *Annotation-Free Reasoning*: it does not require human-curated rationales or verified chain-of-thought. The RL rewards operate on the final answer only (via Reference-based LLM judge and embeddings), significantly lowering data curation cost while still encouraging interpretable reasoning. Collectively, these properties contribute to MediX-R1's strong performance across multiple medical benchmarks and enhance its practical effectiveness for real-world deployment.

For medically relevant performance evaluation, instead of relying on brittle metrics such string-overlap and binary-matching, we employ a ***reference-based LLM-as-judge*** framework that leverages an LLM to assess the semantic correctness and contextual alignment of *model predictions against corresponding question-specific ground-truth answers* through a unified three-stage evaluation process. The evaluations span tasks ranging from short-form QA to long-form report generation, thereby reducing the mismatch between benchmark evaluation and clinical utility. Despite using only ∼51K instruction samples (Fig 1, §A.1), MediX-R1 achieves strong results across diverse medical benchmarks. Our experiments show that the proposed composite rewards not only improve benchmark accuracy, but also mitigate reward hacking and reduce volatility, yielding stable training and interpretable outputs. Compared to open-weight medical models listed in Fig 2, MediX-R1 combines broad modality coverage with single-stage RL and structured reasoning.

Contributions: (i) We introduce *open-ended medical reinforcement learning* by extending Group based RL with tailored rewards for clinical reasoning. **(ii)** We design a *composite reward* with LLM-based accuracy and medical semantic signals that enables open-ended responses with RL in the medical domain and stabilizes training. **(iii)** We employ a three-stage *Reference-based LLM-as-judge evaluation framework* that unifies benchmarking for both LLM (text-only) and VLM (image+text) tasks in the medical setting. **(iv)** MediX-R1 achieves excellent LLM and VLM results with a *single-stage RL recipe using ∼51K instructions*, validated through both Reference-based LLM-as-judge and human expert evaluations. **(v)** Finally, we show the effectiveness of the proposed composite reward across different Group based RL algorithms, achieving consistent performance gains with GRPO (Shao et al., 2024), DAPO (Yu et al., 2025) and GSPO (Zheng et al., 2025a). Moreover, we have conducted experiments on different baseline VLMs, including Qwen2.5-VL, Qwen3-VL (Bai et al., 2025), and SmolVLM2 (Marafioti et al., 2025), and achieved consistent performance gains across these backbones.

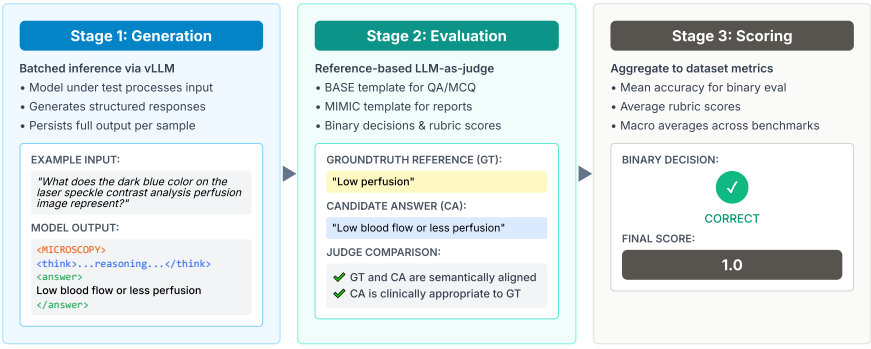

Figure 4: **Evaluation Framework:** Our three-stage evaluation pipeline: (1) Generation via vLLM inference on the model under test, (2) Evaluation using Reference-based LLM-as-judge with BASE and MIMIC templates, and (3) Scoring through aggregation of judgment outputs. The framework supports both binary decisions for QA/MCQ tasks and rubric-based scoring for long-form reports, ensuring robust evaluation across diverse medical benchmarks.

2 Open Ended Medical RL

MediX-R1 fine-tunes a baseline multimodal backbone for open-ended medical reasoning using Group Based RL. Given an image I and question q, the vision encoder produces visual tokens that are fused with text tokens and fed to the LLM policy π_θ. The model generates structured outputs of the form: `[modality tag]<think>free-form clinical reasoning</think><answer>final concise answer</answer>`

2.1 Group-based RL with Composite Rewards

Setup: Given an input $\mathbf{v}$ (image I + question q) drawn from $P(\mathbf{V})$, we sample a group of G candidate completions $\{o_i\}_{i=1}^{G}$ from the frozen behavior policy $\pi_{\theta_{\text{old}}}(\cdot \mid \mathbf{v})$. Each completion receives a scalar reward r_i computed by our *composite reward* (Sec. 2.2). We then compute a standardized *group-relative* advantage:

$$A_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^{G})}{\text{std}(\{r_j\}_{j=1}^{G})}.$$

This removes the need for a learned value function while preserving a stable relative learning signal within group.

GRPO objective: GRPO (Shao et al., 2024) updates π_θ using PPO-style clipping on an importance ratio and a KL regularizer to a fixed ref policy π_{ref}:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\mathbf{v},\,\{o_i\}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\min\Big(\rho_i(\theta)A_i,\; \text{clip}(\rho_i(\theta),\,1-\epsilon,\,1+\epsilon)A_i\Big) - \beta\,\mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\Bigg] \tag{1}$$

where $\rho_i(\theta) = \frac{\pi_\theta(o_i|\mathbf{v})}{\pi_{\theta_{\text{old}}}(o_i|\mathbf{v})}$, and $\epsilon, \beta \geq 0$ control clipping and regularization strength.

Composite reward with different optimizers: In addition to GRPO, we run the *same* composite reward with two recent Group based RL family optimizers, DAPO (Yu et al., 2025) and GSPO (Zheng et al., 2025a)and report their comparison in Table 3.

DAPO (efficiency-focused refinements): DAPO keeps the GRPO/PPO clipped structure but improves token efficiency (as summarized in Yu et al. (2025)): it (i) uses *asymmetric clipping* ("Clip-Higher") with a larger upper bound to avoid prematurely zeroing gradients for rare-but-good tokens, and (ii) averages loss over *all generated tokens* rather than per-sample averaging (reducing gradient dilution for long responses). Compact form:

$$\mathcal{J}_{\text{DAPO}}(\theta) = \mathbb{E}_{\mathbf{v},\,\{o_i\}}\Bigg[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)A_i,\; \text{clip}(r_{i,t}(\theta),\,1-\epsilon_{\text{low}},\,1+\epsilon_{\text{high}})A_i\Big)\Bigg] \tag{2}$$

where $r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|\mathbf{v},o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|\mathbf{v},o_{i,<t})}$. DAPO further encourages rollouts via *dynamic sampling* (avoiding degenerate groups with identical rewards) and can apply *overlong reward shaping* to softly penalize excessively long outputs.

GSPO (sequence-level ratio for stability): GSPO addresses variance from token-wise importance ratios by replacing $r_{i,t}(\theta)$ with a *length-normalized sequence-level ratio* shared across all tokens in a completion. It then uses the same clipped objective form but with $s_i(\theta)$ in place of token-wise ratios:

$$\mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}_{\mathbf{v},\,\{o_i\}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\min\Big(s_i(\theta)A_i,\; \text{clip}(s_i(\theta),\,1-\epsilon,\,1+\epsilon)A_i\Big)\Bigg] \tag{3}$$

Because all tokens within a sequence share the same ratio, GSPO reduces intra-sequence variance and tends to be more stable in settings where token-level ratios are noisy (e.g., long outputs and dynamic architectures).

2.2 Reward Design

We define a composite reward:

$$r = w_{\text{fmt}}\,R_{\text{format}} + w_{\text{llm}}\,R_{\text{llm}} + w_{\text{emb}}\,R_{\text{embed}} + w_{\text{mod}}\,R_{\text{modality}}$$

with coefficients chosen via a simple, staged ablation procedure. In all experiments we keep a small, fixed weight on the format reward to enforce parseable outputs, and distribute the remaining weight across semantic correctness (LLM judge

Benchmarks	MedVLM R1 2B	BiMediX2 8B	Huatuo GPT-V 7B	MedGemma 1.5 4B	MedGemma 27B	MedMO 8B	MediX-R1 2B	MediX-R1 8B	MediX-R1 30B
MMLU-Clinical	0.540	0.732	0.721	0.706	0.879	0.864	0.660	0.845	**0.894**
MMLU-Bio	0.549	0.792	0.708	0.708	0.972	0.951	0.806	0.951	**0.993**
MMLU-Med	0.451	0.694	0.653	0.578	0.866	0.827	0.699	0.879	**0.890**
MMLU-Genetics	0.560	0.790	0.710	0.760	0.940	0.900	0.680	0.900	**0.980**
MMLU-ProfMed	0.500	0.695	0.625	0.768	0.912	0.890	0.628	0.868	**0.974**
MMLU-Anat	0.519	0.659	0.600	0.519	0.793	0.785	0.563	0.763	**0.874**
MedMCQA	0.408	0.572	0.511	0.538	0.727	0.662	0.492	0.683	**0.781**
MedQA	0.400	0.583	0.534	0.690	0.866	0.848	0.497	0.796	**0.929**
USMLE-SA	0.378	0.591	0.538	0.726	0.895	0.742	0.505	0.822	**0.951**
PubMedQA	0.520	0.520	0.542	0.460	0.414	**0.586**	0.472	0.482	0.490
MIMIC-CXR-Sum	0.704	0.672	0.707	0.699	**0.767**	0.709	0.739	0.746	0.765
SLAKE-VQA	0.434	0.468	0.545	0.565	0.634	0.479	0.654	**0.703**	0.683
RadVQA	0.404	0.530	0.614	0.572	0.585	0.419	0.539	0.596	**0.625**
PathVQA	0.239	0.323	0.374	0.332	0.322	0.272	0.428	**0.455**	0.445
PMC-VQA	0.398	0.482	0.532	0.474	0.478	0.360	0.491	0.554	**0.571**
PMC-VQA-Hard	0.020	0.229	0.261	0.258	**0.354**	0.177	0.284	0.317	0.307
MIMIC-CXR-Gen	0.240	0.124	0.316	0.273	0.224	0.084	0.280	0.328	**0.350**
AVG	0.427	0.556	0.558	0.566	0.684	0.621	0.554	0.688	**0.736**

Table 1: **Evaluation Benchmark:** The top section lists LLM (text-only) tasks and the bottom section lists VLM (image+text) tasks. Our three-stage evaluation setting evaluates both tasks in a unified framework. MediX-R1 achieves the highest average score across this diverse suite, exhibiting state-of-the-art performance among open models. Best and second best results are bold and underlined.

/ embeddings) and modality grounding. The coefficient selection procedure is described in Sec. 4.3, and full training config in §A.2.

Why this enables open-ended medical RL : Unlike prior RL setups that are limited to verifiable signals or MCQ-style accuracy (e.g., exact match, executable or rule-based graders), our LLM-based accuracy reward R_{llm} and embedding-based semantic reward R_{embed} provide reliable feedback for free-form, clinically accurate answers. The Reference-based LLM-as-judge converts semantic correctness into a robust YES/NO decision under paraphrase and clinical phrasing, while medical-domain embeddings supply a complementary content-alignment signal. This dual signal makes Reinforcement Learning viable for open-ended medical reasoning; the format (R_{format}) and modality (R_{modality}) rewards act as structural regularizers, but R_{llm} and R_{embed} are the primary drivers of open-ended RL in MediX-R1.

LLM-Based Accuracy Reward (R_{llm}) : We parse the model output's final answer between `<answer>...</answer>` and compare it to the reference answer using a compact Reference-based LLM-as-judge prompt that forces a strict YES/NO decision. Concretely, a local vLLM (Kwon et al., 2023) endpoint (Qwen3-4B) returns YES if the candidate semantically answers the reference, and NO otherwise; we map YES$\mapsto$ 1, NO$\mapsto$ 0. This captures correctness and robustness to paraphrasing while keeping the signal discrete.

Embedding-Based Reward (R_{embed}) : To further encourage semantic alignment, we compute cosine similarity between the predicted answer and the reference using a medical embedding model MedEmbed-large (Balachandran, 2024). We convert it to a binary reward via a threshold (default 0.8): $R_{\text{embed}}{=}1[\cos(\mathbf{e}_{\text{pred}}, \mathbf{e}_{\text{ref}}) \geq \tau]$. This complements the LLM judge and helps capture terminological variants.

Format Reward (R_{format}) : We enforce structured outputs by matching the regex for the exact pattern `<think>...</think> <answer>...</answer>` after normalizing stray whitespace around brackets. Outputs that match receive 1, else 0. This stabilizes training and improves interpretability of the medical reasoning.

Modality recognition Reward (R_{modality}) : We encourage explicit grounding to the imaging modality by requiring the model to emit the predicted modality tag before the `</think>` block (case-insensitive). We compare it to the reference modality tag and assign 1 on match, 0 otherwise. This reduces cross-modality hallucinations (e.g., describing CT findings on an X-ray).

3 Evaluation Framework

Our evaluation pipeline has three stages: Generation, Evaluation, and Scoring. We evaluate across both text-only (LLM) and image+text (VLM) tasks covering QA, MCQ, and long-form report tasks.

Generation. We run batched inference via vLLM on the model under test and persist the full response

per sample. For models that emit structured reasoning, we retain the entire output but, for scoring, discard internal chains-of-thought by stripping content up to and including the closing `</think>` tag, evaluating only the final answer block.

Evaluation: We employ a separate Reference-based LLM-as-judge, Qwen3-14B (Yang et al., 2025), served with vLLM for throughput and stability on modest GPUs. Two prompt families are used: a BASE template (§A.10) for open-ended, one-word, and MCQ-style questions that yields a binary decision, and a MIMIC template (§A.10) for long-form report generation that scores along clinical criteria. For example, on a visual question answering item asking "*What does the dark blue color on the laser speckle contrast analysis perfusion image represent?*" with ground truth "*Low perfusion*" a model response that includes hidden reasoning and the final answer "*Low blood flow or less perfusion*" is judged correct and assigned a score of 1. The judge compares predicted answers against references, accounting for paraphrase and clinically equivalent phrasing. **Scoring:** We aggregate judgment outputs to dataset-level metrics. For binary evaluations, we report mean accuracy over samples. For long-form, we average the scalar rubric scores across samples, optionally normalizing for comparability. We also compute macro averages across benchmarks. **Why Reference-based LLM-as-judge (via vLLM):** Traditional string-overlap metrics (BLEU, ROUGE, F1) for reference (ground truth) comparison often under-reward correct, clinically appropriate paraphrases and cannot assess justification quality or contextual alignment. A Reference-based LLM judge captures semantic correctness, clinical reasoning, and adherence to task-specific criteria through carefully designed prompts, while vLLM serving ensures consistent, fast, and reproducible evaluations.

4 Experiments and Results

4.1 State-of-the-art Comparisons

We evaluate MediX-R1 on a comprehensive suite of medical language and vision-language benchmarks, covering both text-only (LLM) and image+text (VLM) tasks. The evaluation includes standard medical QA, multiple-choice, and open-ended report generation, as well as visual question answering and clinical image interpretation. The datasets used for evaluation are listed in §A.1.

For each dataset, we follow the evaluation as described in the previous section, using Reference-based LLM-as-judge scoring for both short-form and long-form responses. Table 1 summarizes performance on our unified benchmark suite across several open-weight medical models, including MedVLM-R1 *(2B)*, BiMediX2 *(8B)*, HuatuoGPT-V *(7B)*, MedGemma *(4B/27B)*, MedMO *(8B)* (Deria et al., 2026) and MediX-R1 *(2B/8B/30B)*.

Model	MMMU Medical Validation
MedVLM-R1 2B (Pan et al., 2025)	39.33
MedGemma 1.5 4B (Sellergren et al., 2026)	50.00
HuatuoGPT-V 7B (Chen et al., 2024b)	47.33
BiMediX2 8B (Mullappilly et al., 2025)	39.33
Qwen3-VL 8B (Bai et al., 2025)	62.66
MedMO 8B (Deria et al., 2026)	62.66
Qwen3-VL 30B (Bai et al., 2025)	68.66
MedGemma 27B (Sellergren et al., 2025)	56.66
Lingshu 32B (Team et al., 2025)	62.66
MediX-R1 2B	44.66
MediX-R1 8B	64.00
MediX-R1 30B	75.33

Table 2: **MMMU Medical Val results:** Accuracy on the Health & Medical validation subset of MMMU.

MediX-R1 achieves the highest average score across all benchmarks, outperforming prior models on both language and vision-language tasks. Notably, it shows strong gains on open-ended and clinically complex tasks such as report generation, as well as robust performance on standard QA, VQA and crowdsourced real-world benchmarks. These results highlight the effectiveness of our open-ended RL training and composite reward design, which enable MediX-R1 to generate accurate, semantically aligned, and clinically auditable responses beyond the capabilities of models trained only with supervised or MCQ objectives.

We additionally report results on the Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) benchmark (Yue et al., 2024). We select the Health & Medical validation subset (MMMU-Med-Val), covering Basic Medical Science, Clinical Medicine, Diagnostics and Laboratory Medicine, Pharmacy, and Public Health. Results are shown in Table 2.

4.2 Ablation Experiments

Composite Reward across RL algorithms: Our RL-method ablation in Table 3 shows that the proposed composite-reward training transfers across different RL frameworks, consistently improving over the baseline across both LLM and VLM tasks. DAPO achieves the best overall average (0.610), compared to GRPO (0.59), GSPO(0.600), and the baseline Qwen2.5-VL achieves (0.570).

Reward Design Ablation: Table 4 compares reward variants that differ in which non-format signals are active (all settings include the same R_{format}). The Default reward baseline that uses basic string matching against the ground truth is brittle to paraphrases and clinical synonymy. Using only the embedding reward underperforms on text-only evaluations (0.640) and provides gains on VLM (0.409), suggesting that thresholded cosine similarity alone lacks discriminative power for nuanced clinical reasoning. Using only the LLM judge improves text-only accuracy (0.666) but does not help VLM (0.400), indicating that the judge alone is insufficient to enforce modality grounding. All reward design models are compared with checkpoints before reward hacking.

Combining LLM + embedding increases robustness to paraphrase and terminology variants, improving text-only scores (0.686) and yielding a small VLM lift (0.410), which raises the overall average to 0.589. Finally, the full MediX-R1 composite (LLM accuracy + embedding semantics + modality recognition, with shared format control) produces the strongest image+text performance (0.431) while matching the best text-only result (0.687), achieving the best overall average (0.597). Together with Fig. 5, which shows reduced volatility and fewer signs of reward hacking, these results suggest that the composite reward both improves aggregate performance and stabilizes optimization.

Key takeaways: *(i)* The LLM judge is the strongest single signal for text correctness, and embeddings complement it by reducing false negatives from paraphrases.*(ii)* Default string matching degrades substantially on open-ended multimodal evaluations.*(iii)* Modality recognition is important for VLM tasks and drives gains in image+text tasks; the full composite delivers the best overall results.

Performance across VLM backbones: We observe consistent improvements across VLM backbones after training these models in our MediX-R1 framework with our composite rewards as shown in Table 5. These results show that MediX-R1 enhances open-ended medical reasoning ability across backbone models.

Judge Robustness and Evaluator sensitivity. To ensure robustness, we perform controlled evaluations using deterministic generation settings (`temperature=0`, `top_p=1`) and report averages over three runs, observing only ± 0.002 variation. For additional validation, we replaced Qwen3-14B with GPT-5.1 and GPT-5 mini as evaluators, which resulted in a deviation of only ± 0.005, indicating high consistency across judge models.

Evaluations	Baseline	GRPO	GSPO	DAPO
LLM Evaluations (text only)	0.675	0.687	0.689	0.701
VLM Evaluations (image + text)	0.376	0.431	0.439	0.445
Overall AVG	**0.570**	**0.597**	**0.600**	**0.610**

Table 3: **RL Method ablation across benchmarks:** Using the same composite reward, different RL algorithms (GRPO/GSPO/DAPO) consistently outperform the baseline across both tasks, with DAPO achieving the highest overall average.

Evaluations	Default Reward	Embedding Reward	LLM Reward	LLM + Embedding	MediX-R1
LLM Tasks	0.660	0.640	0.666	0.686	0.687
VLM Tasks	0.382	0.409	0.400	0.410	0.431
Overall AVG	**0.562**	**0.558**	**0.572**	**0.589**	**0.597**

Table 4: **Reward ablation on validation benchmarks:** The MediX-R1 composite reward achieves the best overall average, outperforming single-signal and LLM+embedding rewards.

Model	Baseline	+ Composite Rewards
SmolVLM2-2.2B (Marafioti et al., 2025)	0.410	0.432
Qwen3-VL-2B (Bai et al., 2025)	0.529	0.554
Qwen3-VL-8B (Bai et al., 2025)	0.666	0.688
Qwen3-VL-30B (Bai et al., 2025)	0.698	0.736

Table 5: **Baseline comparison across backbones:** Baseline is the original backbone; + Composite Rewards applies our RL with composite rewards on the same backbone.

4.3 Reward Hacking and Mitigation

In reinforcement learning, Reward Hacking occurs when a model maximises its reward in unintended ways, often bypassing the true objective. It arises when the policy exploits imperfections in a single reward signal to earn high scores without producing clinically correct answers. We observed two concrete modes (examples abbreviated):

Embedding model exploit When using Embedding models like MedEmbed-large (Balachandran, 2024) short or non-semantic tokens can spuriously yield high cosine similarity. For instance, a candidate that outputs `<answer>-</answer>` for "What does the white arrow point to in image B?" received $R_{\text{embed}}=1.0$ against the ground truth "Renal artery," despite being incorrect.

LLM judge exploit When using LLMs like Qwen3-4B (Yang et al., 2025) as a rewarder template-like placeholders can confuse the judge when the reference is provided for comparison. E.g., `<answer>`*The largest organ in the picture is [insert your answer here based on the medical reasoning provided above].*`</answer>` was judged correct ($R_{\text{llm}}=1.0$) against the reference "Lung."

Mitigation in MediX-R1 To curb these failures, MediX-R1 employs a composite reward and input/output constraints: *(i) Composite objective:* $R_{\text{llm}} + R_{\text{embed}} + R_{\text{modality}}$ (with shared R_{format}) reduces reliance on any single brittle signal and penalizes mismatches in content or modality recognition (Table 4). *(ii) Embedding gating:* set R_{embed}=0 for answers below a minimum character/word length, with high punctuation or non-alphanumeric ratio; strip punctuation before embedding; calibrate the similarity threshold. *(iii) Modality recognition:* R_{modality} requires a correct modality tag, curbing visually ungrounded shortcuts that might still fool text-only rewards. *(iv) Structural control and regularization:* R_{format} enforces parseable outputs; Group relative advantage and a KL penalty to the reference reduce collapse to degenerate hacks by discouraging outlier behaviors. *(v) Reward coefficient selection (ablation-driven):* To make the reward design transparent, we selected coefficients via a staged procedure rather than an exhaustive hyperparameter search. Concretely, we fixed w_{fmt}=0.10 in all experiments and allocated the remaining 0.90 mass to the task-facing signals. This yields the ablation settings in Table 4, e.g., embedding-only $r = 0.1R_{\text{format}} + 0.9R_{\text{embed}}$ and LLM-only $r = 0.1R_{\text{format}} + 0.9R_{\text{llm}}$. For the combined semantic reward, we evaluated a small set of intuitive splits between R_{llm} and R_{embed} while keeping w_{fmt} fixed; performance was similar across these choices on our validation benchmark, and we selected the configuration that slightly favored the LLM judge. Finally, when adding modality grounding, we reserved 5% of the non-format budget for R_{modality}, and renormalized the remaining non-format weights. (See §A.3 for full coefficient-selection details, including all $R_{\text{llm}}/R_{\text{embed}}$ splits.)

Together, these measures reduce reward hacking and stabilize training, yielding smoother rewards and stronger final performance (Fig. 5).

4.4 Human Expert Evaluation

To assess the clinical quality of model outputs, we conducted a human expert evaluation using a blind review setup (See Evaluation Protocol in §A.6). For a randomly selected subset of questions from our Evaluation benchmark, responses were generated by four models: MediX-R1, Llama3.2-Vision, MedGemma and HuatuoGPT-Vision. The outputs were anonymized and labeled as Model A, Model B, Model C and Model D with no identifiers provided to the reviewers. Medical experts were asked to evaluate the responses against the provided ground truth descriptions for each question. The evaluation focused on determining which model produced the most accurate, clinically relevant response along with interpretable reasoning traces.

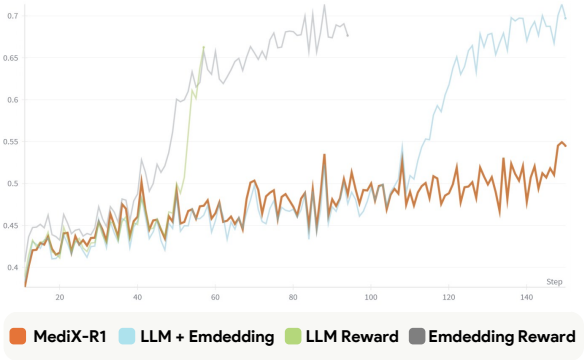

Figure 5: **Overall validation reward vs training step across reward designs**. Training with individual signals LLM-only or embedding-only shows volatility and reward hacking, while LLM+embedding reduces but does not eliminate instability. MediX-R1 uses a composite reward which stabilizes learning and delivers the highest final reward and accuracy.

The results shows a strong preference for MediX-R1, which was selected as the best response in 72.7% of the cases. In comparison, Llama3.2-Vision was preferred in 13.6% of the cases, MedGemma in 9.2% and HuatuoGPT-Vision in 4.5% of the cases. Additional details on human expert evaluation is available in Sec. §A.6 and Sec. §A.7. Furthermore, we provide additional results on crowd-sourced clinical data (MedPix 2.0) in § A.8, alongside detailed qualitative examples illustrating MediX-R1's clinically auditable reasoning outputs across modalities in § A.5 (see Fig. 6).

5 Conclusion

We presented MediX-R1, an open-ended reinforcement learning framework for medical multimodal reasoning. By training a baseline VLM with Group-based RL and a composite reward that combines a reference-based LLM judge, medical embedding-based semantic alignment, format control, and modality recognition, MediX-R1 learns to generate accurate responses with interpretable reasoning traces. The model achieves strong results across multimodal medical benchmarks and human evaluation while improving training stability and resistance to reward hacking. Together, our multi-signal reward design and standardized evaluation provide a robust, scalable path for interpretable medical multimodal RL fine-tuning.

6 Limitations

MediX-R1 is a research prototype and is not validated for clinical deployment. Although the proposed composite reward improves open-ended medical reasoning, the model may still hallucinate findings, omit important clinical details or key differentials, and produce confident but incorrect answers. Therefore, MediX-R1 should not be used for autonomous diagnosis, treatment recommendation, triage, or clinical decision-making.

While MediX-R1 produces structured reasoning traces, these should not be interpreted as guaranteed faithful explanations of the model's decision process. In this work, we enforce the presence of reasoning traces through a format reward, but we do not explicitly validate the clinical correctness or faithfulness of the reasoning traces with a dedicated reward function. Instead, our human expert evaluation compares the overall quality of MediX-R1's responses, including reasoning, against other medical models. Developing explicit reward functions and evaluation protocols for faithful medical reasoning remains a promising direction for future work.

Our reward and evaluation pipeline relies on Reference-based LLM-as-judge and embedding-based signals, which are imperfect proxies for clinical correctness. These signals may reinforce subtle biases, accept false positives, or be sensitive to ambiguity and paraphrasing. Although the composite reward mitigates reward hacking by combining multiple signals, it does not completely eliminate this risk.

The evaluation is limited by the coverage of available benchmarks and public datasets. Results may not generalize to all clinical settings, demographic groups, imaging conditions, rare diseases, languages, or real-world workflows. Moreover, while we use publicly available, de-identified datasets under their respective licenses, residual privacy and data-governance risks may remain. Larger expert evaluations, fairness analyses where legally and ethically permissible, uncertainty calibration, bias-aware reward shaping, clinician-in-the-loop evaluation, and prospective clinical studies are needed before any practical use.

7 Ethical Considerations

This work follows the ACL Ethics Policy and is intended for research and educational use only. MediX-R1 is not validated for clinical or commercial deployment and should not be used for autonomous diagnosis, treatment recommendation, triage, or clinical decision-making without qualified medical oversight.

The main ethical risks are patient harm from incorrect or misleading medical outputs, over-reliance by users, and possible misuse for self-diagnosis, misleading medical narratives, or adversarial prompting. Like other medical generative models, MediX-R1 may also amplify demographic, institutional, or dataset biases, potentially contributing to health disparities. Although we use publicly available, de-identified datasets under their respective licenses, residual privacy and data-governance risks may remain. Downstream users should therefore conduct safety, privacy, fairness, uncertainty, and clinician-in-the-loop evaluations before applying the model in any medical context.

To support responsible use and scrutiny, we commit to releasing training and inference code, configurations, checkpoints, curated datasets, and RL/evaluation prompt templates, alongside a model card and clear usage restrictions, under a CC-BY-NC-SA 4.0 license. We also disclose our use of generative AI tools: assisted coding was limited to boilerplate scaffolding and minor refactors, with all algorithmic logic authored and reviewed manually; writing-support models were used for grammar and style, while all technical claims and results were contributed and verified by the authors. These steps aim to ensure transparency, auditability, and reliable reproduction of the published results.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Abhinand Balachandran. 2024. Medembed: Medical-focused embedding models. https://github.com/abhinand5/MedEmbed. Software.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024a. Huatuogpt-o1, towards medical complex reasoning with llms. *Preprint*, arXiv:2412.18925.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and Benyou Wang. 2024b. Towards injecting medical visual knowledge

into multimodal LLMs at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7346–7370, Miami, Florida, USA. Association for Computational Linguistics.

Ankan Deria, Komal Kumar, Adinath Madhavrao Dukre, Eran Segal, Salman Khan, and Imran Razzak. 2026. Medmo: Grounding and understanding multimodal large language model for medical images. CVPR 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *Preprint*, arXiv:2003.10286.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Alex Henigman and Brandon Kennedy. 2025. Medpix®: database of medical images, teaching cases, and clinical topics. *Medical Reference Services Quarterly*, 44(3):328–333.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2025. Smolvlm: Redefining small and efficient multimodal models. *Preprint*, arXiv:2504.05299.

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Sara Pieri, Saeed Yahya Alseiari, Shanavas Cholakkal, Khaled M Aldahmani, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2025. BiMediX2 : Bio-medical EXpert LMM for diverse medical modalities. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14051–14071, Suzhou, China. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–347. Springer.

Sara Pieri, Sahal Shaji Mullappilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2024. BiMediX: Bilingual medical mixture of experts LLM. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16984–17002, Miami, Florida, USA. Association for Computational Linguistics.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse,

Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, Liron Yatziv, Tiffany Chen, Bram Sterling, Kenneth Philbrick, Richa Tiwari, Yun Liu, Madhuram Jajoo, Chandrashekar Sankarapu, Swapnil Vispute, and 23 others. 2026. Medgemma 1.5 technical report. *Preprint*, arXiv:2604.05081.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Aditya Shourya, Michel Dumontier, and Chang Sun. 2025. Adapting lightweight vision language models for radiological visual question answering. *Preprint*, arXiv:2506.14451.

Irene Siragusa, Salvatore Contino, Massimo La Ciura, Rosario Alicata, and Roberto Pirrone. 2025. Medpix 2.0: A comprehensive multimodal biomedical data set for advanced ai applications with retrieval augmented generation and knowledge graphs. *Data Science and Engineering*, pages 1–17.

LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *Preprint*, arXiv:2506.07044.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Pmc-vqa: Visual instruction tuning for medical visual question answering. *Preprint*, arXiv:2305.10415.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025a. Group sequence policy optimization. *Preprint*, arXiv:2507.18071.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. 2025b. Easyr1: An efficient, scalable, multi-modality rl training framework. `https://github.com/hiyouga/EasyR1`.

A Appendix

A.1 Data and Benchmarks

Training: We trained MediX-R1 on 51335 multimodal medical instruction samples spanning 16 modality tags. All samples were drawn from the official train splits of the source datasets: PMC-VQA subset (Zhang et al., 2024), SLAKE (Liu et al., 2021), RadVQA (Lau et al., 2018), and PathVQA (He et al., 2020).

Evaluation: For evaluation, we used 19,200 samples across LLM and VLM benchmarks from their respective test sets.

LLM (text-only) benchmarks: MMLU-Clinical, MMLU - Bio, MMLU - Med, MMLU - Genetics, MMLU - ProfMed, MMLU - Anat (Hendrycks et al., 2020), MedMCQA (Pal et al., 2022), MedQA (Jin et al., 2021), USMLE - SA (Han et al., 2023), PubMedQA (Jin et al., 2019), MIMIC - CXR - Summarization (Johnson et al., 2016).

VLM (image+text) benchmarks: SLAKE-VQA (Liu et al., 2021), RadVQA (Lau et al., 2018), PathVQA (He et al., 2020), PMC-VQA (Zhang et al., 2024), PMC-VQA-Hard, MIMIC-CXR-Report Generation (Johnson et al., 2019).

Table 8 summarizes the evaluation benchmark composition, with the upper section listing LLM tasks and the lower section listing VLM tasks.

Medical Modality	Samples
X_RAY	5964
MICROSCOPY	16399
CLINICAL_PHOTOGRAPHY	8979
CT_SCAN	7646
GRAPHICS	2205
ANGIOGRAPHY	522
PET_SCAN	406
ULTRASOUND	1227
MRI_SCAN	6224
FUNDUS_PHOTOGRAPHY	314
OCT_SCAN	236
ENDOSCOPY	611
MAMMOGRAPHY	106
FLUOROSCOPY	321
OTHER	64
SPECT	111
Total	**51335**

Table 6: **Modality Breakdown.**

A.2 Training Configuration

The GRPO training configuration used for MediX-R1 is given in Listing A.2. All training configurations and hyperparemeters and available in the *anonymous codebase*. Core settings include (i) data filtering and batching, (ii) actor optimization and rollout sampling, (iii) KL-regularized GRPO advantage computation, and (iv) trainer settings. We train our models using the EasyR1(Zheng et al., 2025b) Github Repository. MediX-R1 was trained using 8×A100 (80 GB) Nvidia GPUs for approximately 25 hours.

Dataset	Samples
PMC_VQA_SUBSET	25000
SLAKE	4919
RAD_VQA	1793
PATH	19623
Total	**51335**

Table 7: **Source Dataset Composition.**

A.3 Reward Coefficient Selection Details

This section details how we selected the composite reward coefficients used throughout the paper. Our goal was to (i) keep outputs reliably parseable (format control), while (ii) allocating most of the reward budget to task-facing correctness signals, and (iii) avoiding an expensive hyperparameter search.

Fixed format budget. Across all experiments (including Table 4), we fixed the format reward weight to $w_{\text{fmt}} = 0.10$ to enforce consistently parseable outputs. The remaining 0.90 of the reward mass was allocated among the task-facing signals.

Single-signal ablations. The embedding-only and LLM-only settings in Table 4 correspond to:

$$r_{\text{emb-only}} = 0.1R_{\text{format}} + 0.9R_{\text{embed}}$$
$$r_{\text{llm-only}} = 0.1R_{\text{format}} + 0.9R_{\text{llm}}$$

Combining semantic rewards (R_{llm} + R_{embed}). Next, we combined the LLM judge and embedding signals while keeping $w_{\text{fmt}} = 0.10$ fixed, and evaluated three intuitive splits of the remaining 0.90 reward contribution:

v1 (equal split):

$$r = 0.1R_{\text{format}} + (0.5 * 0.9)R_{\text{llm}} + (0.5 * 0.9)R_{\text{embed}}$$

v2 (LLM-favored):

$$r = 0.1R_{\text{format}} + (0.6 * 0.9)R_{\text{llm}} + (0.4 * 0.9)R_{\text{embed}}$$

v3 (embed-favored):

$$r = 0.1R_{\text{format}} + (0.4 * 0.9)R_{\text{llm}} + (0.6 * 0.9)R_{\text{embed}}$$

On our validation benchmark suite, these three variants yielded similar overall averages (v1: 0.582,

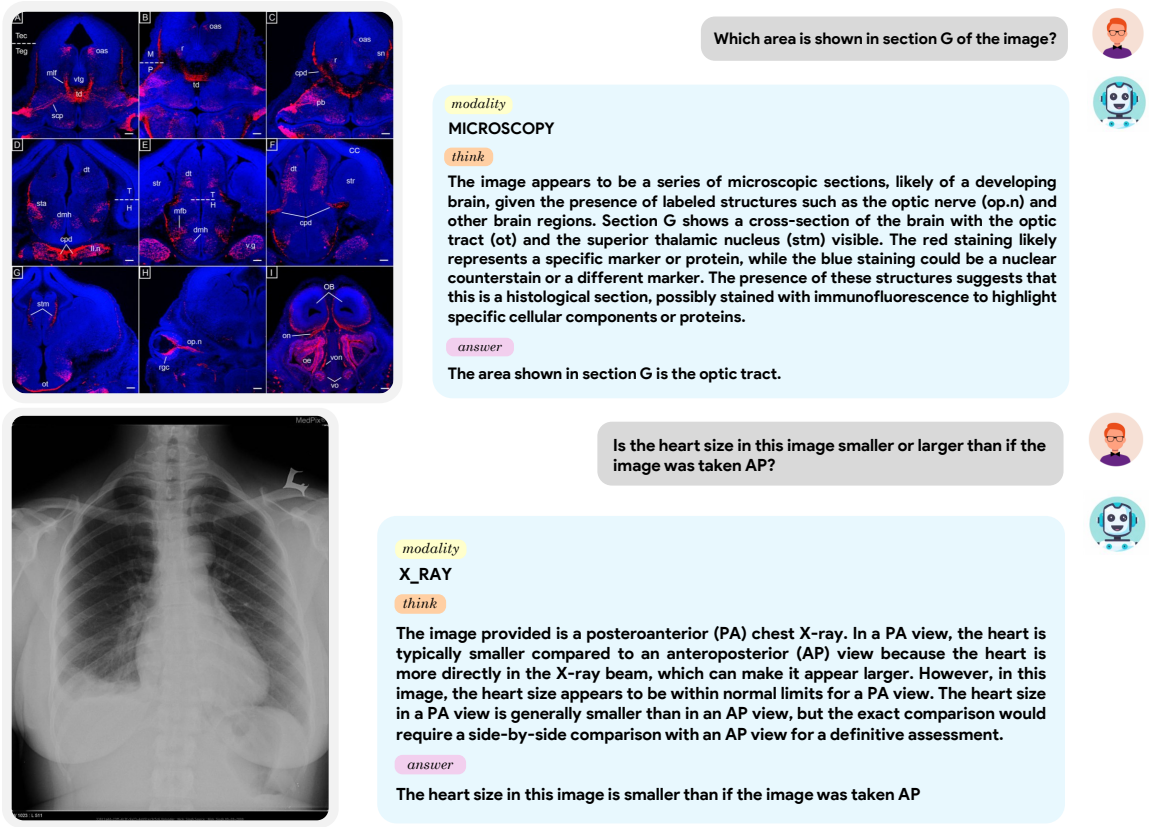

Figure 6: **Qualitative examples of MediX-R1**. (Top, Microscopy) Correctly identifies the optic tract in section G with interpretable reasoning. (Bottom, X-ray) Explains why heart size appears smaller in PA vs. AP view. MediX-R1 generates clinically auditable, open-ended answers across modalities.

v2: 0.589, v3: 0.579), indicating low sensitivity within this coarse range. We therefore selected **v2** as the default combined-semantic configuration because it slightly improved aggregate performance while placing more weight on the stricter correctness signal (R_{llm}).

Adding modality grounding. Finally, we introduced modality grounding by reserving 5% of the *non-format* budget for R_{modality} (i.e., $0.05 \times 0.90 = 0.045$ of total reward mass), and renormalizing the remaining non-format weights according to v2:

$$w_{\text{fmt}} = 0.10 \qquad w_{\text{mod}} = 0.045$$
$$w_{\text{llm}} = 0.5175 \qquad w_{\text{emb}} = 0.3375$$

which sums to 1.0 and matches the MediX-R1 composite setting used in Table 4.

Compute limitations. We emphasize that we did not run an exhaustive hyperparameter search over reward coefficients due to computational constraints. The above staged procedure was intended to make coefficient selection transparent and reproducible; future work may further improve performance by exploring a broader coefficient grid.

A.4 Reward Function Source Code

The Python implementations of the four reward components used in MediX-R1 are shown in Listing A.4 (Format reward), Listing A.4 (LLM-based accuracy reward), Listing A.4 (Embedding-based semantic reward), and Listing A.4 (Modality recognition reward). Each function operates on a predicted model output string and a ground truth string containing the modality tag and reference answer.

A.5 Qualitative Examples

Fig. 6 illustrates how MediX-R1's structured outputs and composite reward translate into clinically auditable behavior across modalities. *Microscopy (top)* Given a multi-panel histological image and the question "Which area is shown in section G of the image?," the model (i) correctly emits the modality tag (`MICROSCOPY`), (ii) provides interpretable reasoning inside `<think>` that references recognizable neuroanatomical markers (e.g., optic tract "ot," superior thalamic nucleus "stm"), stain patterns, and panel context, and (iii) produces a concise final answer: "the optic tract." The modality recognition and format rewards ensure the answer

Benchmark	Samples
mmlu_clinical_knowledge	265
mmlu_college_biology	144
mmlu_college_medicine	173
mmlu_medical_genetics	100
mmlu_professional_medicine	272
mmlu_anatomy	135
medmcqa	4183
medqa	1273
usmle_sa	325
pubmedqa	500
mimic_cxr_report_summarization	100
slake_vqa	1061
rad_vqa	451
path_vqa	6719
pmc_vqa	1997
pmc_vqa_hard	1402
mimic_cxr_report_generation	100
Total	**19200**

Table 8: **Evaluation Benchmark:** The top section lists LLM tasks (text-only), while the bottom section lists VLM (image+text) tasks.

Model	Score (%)
MedMO 8B	33.91
MedVLM-R1	37.24
LLaVA-Med 7B	44.29
BiMediX2 8B	46.51
HuatuoGPT-V 7B	48.81
MedGemma 1.5 4B	47.80
MedGemma 27B	57.88
MediX-R1 2B	**44.01**
MediX-R1 8B	**56.24**
MediX-R1 30B	**60.78**

Table 9: Performance comparison on the MedPix 2.0 dataset. (Crowdsourced Clinical Cases)

is localized to the requested panel and presented cleanly in the <answer> block, while the LLM and embedding rewards bias the policy toward semantically correct identification despite diverse phrasing in the reasoning. *X-ray (bottom)* For "Is the heart size in this image smaller or larger than if the image was taken AP?," the model tags the modality as X_RAY and reasons about projection geometry: PA views reduce cardiac magnification relative to AP due to a shorter heart-to-detector distance and standard source-to-image distance. The model explains this in <think> and answers "smaller" in <answer>. This example shows the model using domain knowledge rather than superficial pattern matching, with the final answer isolated for scoring. Additionally, Figures 7 and 8 show MediX-R1's comprehensive reasoning capabilities for full medical report generation across two different cases.

A.6 Human Expert Evaluation Protocol

As outlined in Listing A.6, for a sampled set of multimodal questions, four anonymized model outputs (A-D) plus a reference description are shown; experts pick the single best response based on clinical correctness, relevance (no hallucinations), and clarity of reasoning. Votes are aggregated into preference percentages reported in the main text.

A.7 Human Evaluation: Model Reasoning

We extend our human expert study detailed in (Sec. 4.4) to evaluate the reasoning quality of our MediX-R1 model against MedGemma with the help of medical doctors. Experts assessed outputs for clinical accuracy, reasoning soundness, and practical usefulness in a medical setting. MediX-R1's reasoning was preferred in 74.2% of cases over MedGemma, indicating stronger clinical coherence. Furthermore, the study shows that in 92.4% of the cases, the model's reasoning steps were rated as acceptable and often comparable to a medical doctor's thought process, while only 7.6% of the cases were rated as having poor reasoning quality. Moreover, in fewer than 5% of the cases, the model produced flawed reasoning despite generating the correct final answer, indicating that such inconsistencies are rare and that MediX-R1 generally maintains a robust and coherent reasoning process. Reviewers comprised five certified medical experts (MBBS/MD) with specialties in Radiology, General Medicine, and Forensic Medicine, with an inter-rater agreement of 63%.

A.8 Evaluation on Real World Clinical Data

To further assess the generalization ability of our model, we conducted additional evaluation on MedPix 2.0 (Siragusa et al., 2025), a publicly available real-world clinical VQA dataset derived from the original MedPix (Henigman and Kennedy, 2025) database maintained by the U.S. National Library of Medicine (NIH). MedPix comprises over 12,000 anonymized, crowdsourced clinical cases containing medical images and corresponding textual information such as findings, diagnoses, and treatments. This ensures both reproducibility and compliance with NIH privacy standards.

The evaluation on MedPix 2.0 shows that MediX-R1 consistently performs strongly across model scales and achieves the best overall result among the evaluated medical vision-language mod-

els (Shourya et al., 2025). Specifically, MediX-R1 30B achieves a score of 60.78%, outperforming strong baselines and previous state-of-the-art medical models, as shown in Table 9. These results further confirm the robustness, scalability, and adaptability of MediX-R1 on diverse real-world clinical data, emphasizing its capability to generalize beyond controlled experimental environments.

A.9 Reinforcement Learning Training Prompt

As shown in Listing A.9, the RL training prompt enforces (i) an explicit modality tag, (ii) structured reasoning in `<think>...</think>`, and (iii) a concise final answer in `<answer>...</answer>`. These structures align with the format reward (R_{format}) and modality reward (R_{modality}) in our composite objective. During training, only the `<answer>` block is graded by the Reference-based LLM-as-judge (R_{llm}) and the embedding-based semantic reward (R_{embed}); the `<think>` content is ignored for scoring but improves interpretability.

Key points:

- Modality tag must be one of the fixed set and appear before `<think>`.
- The final decision is evaluated solely from `<answer>` for R_{llm} and R_{embed}.
- Structural compliance (tags present and ordered) is required for R_{format}.

A.10 Evaluation Templates

We employ two tailored evaluation templates depending on the task format.

Base Prompt: For short-form QA and MCQ-style tasks, we use the BASE evaluation prompt shown in Listing A.10. This judge prompt yields a binary score (0/1) by comparing the predicted `<answer>` against the reference, allowing paraphrases and option-label matches. We use deterministic settings (e.g., temperature=0.0, top_p=1.0) for reproducibility and parse the returned JSON strictly.

Report Generation Prompt: For long-form outputs (e.g., report generation or summarization), we use the report generation evaluation prompt shown in Listing A.10. The judge assigns a rubric score in [0, 5] reflecting clinical accuracy, completeness, and relevance. We request strict JSON for reliable parsing and average scores across items for dataset-level metrics. Only the model's final report text is provided to the judge; any reasoning (e.g., within `</think>`) is stripped before evaluation.

Generate a detailed report based on the scan

think

The image appears to be a chest radiograph (X-ray), likely taken in a posterior-anterior (PA) view, as suggested by the positioning of the heart and spine. The patient seems to be upright during the examination, with the diaphragm in a normal position. The bony structures, including the ribs, clavicles, scapulae, and vertebrae, are clearly visible and show no evidence of fractures, dislocations, or deformities. The lung fields appear clear, with no signs of consolidation, masses, or abnormal opacities. The lung markings are normal, and there is no visible pleural effusion or pneumothorax. The heart size is within normal limits, and the mediastinum appears central without widening or abnormal contours. The diaphragm and costophrenic angles are sharp and well-defined, indicating normal lung expansion and no fluid accumulation

The image provided appears to be a radiograph (X-ray) of the chest. Here is a detailed analysis based on the visible features:

General Observations:
- The image is a posterior-anterior (PA) view of the chest, as indicated by the orientation of the spine and the position of the heart shadow.
- The patient is likely standing or sitting upright during the scan, as the diaphragm appears to be in a normal position.

Key Features:
1. Bony Structures:
  - The ribs are visible on both sides, and the clavicles are also visible.
  - The spine is well-defined, with the vertebrae appearing intact and without any obvious fractures or dislocations.
  - The scapulae are visible on both sides, and the shoulder joints appear to be in a normal position.

2. Lungs:
  - The lung fields appear to be clear without any visible opacities, masses, or infiltrates.
  - The lung markings are visible and appear to be within normal limits.
  - There is no evidence of pleural effusion or pneumothorax.

3. Heart and Mediastinum:
  - The heart size appears to be within normal limits.
  - The mediastinum is not widened, and there are no visible masses or abnormalities in this region.

4. Diaphragm:
  - The diaphragm appears to be intact and in a normal position.
  - The costophrenic angles (where the diaphragm meets the ribs) appear sharp and clear, indicating no pleural effusion.

Potential Concerns:
- There are no obvious fractures or dislocations in the bony structures.
- The lung fields appear clear, and there are no visible masses or infiltrates.
- The heart size and mediastinum appear normal.
- The diaphragm is intact and in a normal position.

Conclusion:
Based on the radiographic findings, this chest X-ray appears to be within normal limits. There are no immediate concerns noted in this image. However, a definitive interpretation should be made by a qualified radiologist who can consider the patient's clinical history and symptoms.

If there are any specific symptoms or concerns, it is recommended to consult a healthcare provider for further evaluation.

Figure 7: **MediX-R1 - Report Generation: Case 1**

Training Configuration

```
Training Configurations
  "data": {
    "max_prompt_length": 4352,
    "max_response_length": 4096,
    "rollout_batch_size": 512,
    "val_batch_size": 1024,
    "shuffle": true,
    "seed": 1,
    "min_pixels": 262144,
    "max_pixels": 4194304,
    "filter_overlong_prompts": true,
    "filter_overlong_prompts_workers": 16
  },
  "worker": {
    "hybrid_engine": true,
    "actor": {
      "strategy": "fsdp",
      "global_batch_size": 128,
      "micro_batch_size_per_device_for_update": 1,
      "micro_batch_size_per_device_for_experience": 2,
      "max_grad_norm": 1.0,
      "clip_ratio_low": 0.2,
      "clip_ratio_high": 0.3,
      "clip_ratio_dual": 3.0,
      "loss_avg_mode": "token",
      "padding_free": true,
      "dynamic_batching": true,
      "use_torch_compile": true,
      "optim": {
        "lr": 1e-6,
        "betas": [0.9, 0.999],
        "weight_decay": 0.01,
        "strategy": "adamw",
        "lr_scheduler_type": "constant",
        "training_steps": 200
      },
      "fsdp": {
        "enable_full_shard": true,
        "enable_rank0_init": true,
        "mp_param_dtype": "bf16",
        "mp_reduce_dtype": "fp32",
        "mp_buffer_dtype": "fp32"
      },
      "offload": {
        "offload_params": true,
        "offload_optimizer": true
      },
      "use_kl_loss": true,
      "kl_penalty": "low_var_kl",
      "kl_coef": 0.01
    },
    "rollout": {
      "name": "vllm",
      "n": 5,
      "temperature": 1.0,
      "top_p": 1.0,
      "seed": 1,
      "tensor_parallel_size": 2,
      "max_num_batched_tokens": 8448,
      "gpu_memory_utilization": 0.6,
      "val_override_config": {
        "temperature": 0.6,
        "top_p": 0.95,
        "n": 1
      },
      "prompt_length": 4352,
```

```
    "response_length": 4096
  }
},
"algorithm": {
  "adv_estimator": "grpo",
  "gamma": 1.0,
  "lam": 1.0,
  "use_kl_loss": true,
  "kl_penalty": "low_var_kl",
  "kl_coef": 0.01,
  "kl_type": "fixed",
  "kl_target": 0.1,
  "kl_horizon": 10000.0
},
"trainer": {
  "total_epochs": 2,
  "nnodes": 1,
  "n_gpus_per_node": 8,
  "val_freq": 5,
  "val_before_train": true,
  "save_freq": 5,
  "save_limit": 3
}
```

Format reward

```
def format_reward(predict: str) -> float:
    idx = predict.find("<think>")
    if idx == -1:
        return 0.0
    predict_new = predict[idx:].strip()
    pattern = re.compile(r"<think>.*?</think>\s*<answer>.*?</answer>", re.DOTALL)
    format_match = re.fullmatch(pattern, predict_new)
    return 1.0 if format_match else 0.0
```

LLM-based accuracy reward

```
def accuracy_reward_llm(predict: str, ground_truth: str) -> float:
    try:
        content_match = re.search(r"<answer>(.*?)</answer>", predict, re.DOTALL)
        given_answer = content_match.group(1).strip() \
                       if content_match else predict.strip()
        given_answer = given_answer.strip('.')
        ground_truth = ground_truth.split('>', maxsplit=1)[1].strip()
        ground_truth = ground_truth.strip('.')

        if given_answer == '' or len(given_answer) == 1:
            return 0.0
        if given_answer == ground_truth:
            return 1.0
        llm_score = llm_answer_match(given_answer, ground_truth)
        return llm_score
    except Exception:
        return 0.0
```

Embedding-based semantic reward

```
def accuracy_reward_embed(predict: str, ground_truth: str,\
                          threshold: float = 0.8) -> float:
    try:
        content_match = re.search(r"<answer>(.*?)</answer>", predict, re.DOTALL)
        given_answer = content_match.group(1).strip()
                       if content_match else predict.strip()
        given_answer = given_answer.strip('.')
        ground_truth = ground_truth.split('>', maxsplit=1)[1].strip()
        ground_truth = ground_truth.strip('.')

        if given_answer == '' or len(given_answer) == 1:
            return 0.0
        if given_answer == ground_truth:
            return 1.0

        embeddings = embed_model.encode([given_answer, ground_truth],\
                                        convert_to_tensor=True)
        similarity = util.pytorch_cos_sim(embeddings[0], embeddings[1]).item()
        return float(similarity >= threshold)
    except Exception:
        return 0.0
```

Modality recognition reward

```
def modality_reward(predict: str, ground_truth: str) -> float:
    idx = predict.find("<think>")
    if idx == -1:
        return 0.0
    predict_new = predict[:idx].strip()  # modality tag before <think>
    modality = ground_truth.split('>', maxsplit=1)[0] + '>'
    return 1.0 if predict_new.upper() == modality.upper() else 0.0
```

Evaluation Protocol for Medical Experts

```
Instructions for Evaluation
Your task is to evaluate the responses provided by three AI models based on a given medical
    image description (Ground Truth). Follow these steps to make your selection:

1) Read the Ground Truth: Carefully review the provided description of the medical image.
    This serves as the reference for an accurate and detailed response.

2) Assess the Model Responses: Examine the three model-generated responses (Model A, Model B,
    and Model C, Model D). Compare their content with the Ground Truth, focusing on the
    accuracy, completeness, and relevance of the clinical reasoning

3) Select the Best Response: Choose the model response that best aligns with the Ground Truth
    in terms of:

 > Clinical Accuracy: Does the response correctly describe the key findings in the image?

 > Reasoning Traces: Does the models reasoning traces correct and well explained

4) Submit Your Choice: After evaluating the responses, select the one that provides the most
    accurate and comprehensive explanation.
```

Reinforcement Learning Training Prompt

```
You are a Medical AI Assistant with advanced reasoning capabilities
Your task:
1. First output the image modality tag from this set:
   <X_RAY>, <MICROSCOPY>, <CLINICAL_PHOTOGRAPHY>, <CT_SCAN>, <GRAPHICS>,
   <ANGIOGRAPHY>, <PET_SCAN>, <ULTRASOUND>, <MRI_SCAN>, <FUNDUS_PHOTOGRAPHY>,
   <OCT_SCAN>, <ENDOSCOPY>, <MAMMOGRAPHY>, <FLUOROSCOPY>, <OTHER>, <SPECT>
   (Only output the tag, nothing else.)
2. Then output the thinking and medical reasoning process in <think>...</think> tags.
3. Finally, provide the correct answer inside <answer>...</answer> tags.
4. Do not include any extra information or text outside of these tags.
Question:
<image>{{ content | trim }}
```

Evaluation BASE template Prompt

```
You are a medical expert.

Your task is to evaluate whether the Predicted Answer correctly answers the Medical Question,
    based on the Ground Truth (Correct Answer) provided.

Question:
{question}

Correct Answer:
{correct_answer}

Predicted Answer:
{predicted_answer}

Score 1 if the predicted answer matches the correct answer either fully in text or by
    indicating the correct option label (e.g., "B", "Option B", or a paraphrased version that
    clearly identifies the correct choice). Score 0 if the predicted answer is incorrect or
    points to the wrong option.

Respond strictly in the following JSON format:

'''json
{{
"score": <score>
}}
'''
```

Evaluation Prompt for Report Generation

```
You are a medical expert evaluating the clinical accuracy, completeness, and relevance of a
    generated medical report or summary.

Your task is to compare an AI-generated report or summary to a reference (gold standard)
    report or summary, based on a clinical instruction or question. Assess the generated
    output on how well it preserves key clinical information, factual correctness, and
    clinical reasoning relevant to the task.

Assign a score between 0 and 5 using the following scale:

0 - Completely incorrect: Clinically irrelevant, misleading, or factually wrong. No
    meaningful alignment with the instruction or reference.

1 - Poor match: Barely relevant or mostly incorrect. Contains significant clinical
    misinformation or omits nearly all critical details.

2 - Weak match: Some fragments of relevant content are present, but major clinical errors or
    omissions exist. Clinical utility is low.

3 - Fair match: Contains several relevant points, but includes notable errors, missing
    findings, or misinterpretations that affect clinical reliability.

4 - Good match: Mostly accurate and clinically sound. Minor issues or missing details, but
    the overall meaning and purpose are preserved.

5 - Perfect or near-perfect match: Clinically accurate, complete, and faithful to the
    instruction and reference. No significant omissions or errors.

Respond only in the following example JSON format:

Example JSON format:
'''json
{{
"score": <score between 0 and 5>
}}
'''

Now, evaluate the following:

### Clinical Instruction or Question::
{question}

### Reference Report or Summary:
{correct_answer}

### AI-Generated Report or Summary:
{predicted_answer}
```

Generate a detailed report based on the scan

think

The image is a radiograph, specifically a chest X-ray, which is used to assess the lungs, heart, and mediastinum. The view appears to be posteroanterior (PA), as indicated by the relative heart size and orientation of the structures. The diaphragm is in a normal position, suggesting the patient was upright during the scan.

On inspection, there is increased opacification noted in the right lung field, which may represent an abnormal finding. This could be due to several potential causes such as infection (pneumonia), pulmonary edema, mass lesion, or other infiltrative processes. The left lung field appears relatively clear in comparison. The heart size and mediastinum are within normal limits, showing no evidence of enlargement or shift. The diaphragm is intact with clear costophrenic angles, indicating no pleural effusion.

The bony structures, including the ribs, clavicles, and scapulae, are visible and intact without evidence of fracture or deformity. Overall, the primary observation is the increased opacity in the right lung, which warrants further clinical correlation and radiological evaluation to determine the underlying cause. A detailed review by a qualified radiologist is recommended for confirmation and diagnosis.

The image provided is a chest X-ray. Here is a detailed report based on the visible findings:

General Observations:
- The X-ray is a posteroanterior (PA) view, as indicated by the orientation of the heart and the position of the diaphragm.
- The patient's left side is on the right side of the image, and the right side is on the left side of the image, which is typical for a PA view.

Lung Fields:
- The lung fields appear to have increased opacification, particularly in the right lung field. This increased opacification could be due to several reasons, including but not limited to:
  - Infiltration (e.g., pneumonia, pulmonary edema, or consolidation).
  - Mass lesion.
  - Pleural effusion.
  - Other pathologies such as fibrosis or tumor.
- The left lung field appears relatively clear compared to the right lung field.

Heart:
- The heart size appears to be within normal limits on this view.
- The mediastinum appears to be within normal limits.

Diaphragm:
- The diaphragm appears to be intact and without any obvious abnormalities.
- The costophrenic angles appear clear, which is a normal finding.

Bones:
- The bony structures of the thorax, including the ribs, clavicles, and scapulae, appear intact without any obvious fractures or lytic lesions.

Pleural Spaces:
- There appears to be no evidence of pleural effusion on this view.

Summary:
The X-ray shows increased opacification in the right lung field, which could be indicative of a variety of conditions. Further clinical correlation with the patient's history, symptoms, and other diagnostic tests (such as a CT scan or sputum analysis) would be necessary to determine the exact cause of the opacification. A radiologist's interpretation is recommended for a definitive diagnosis.

Recommendations:
- The patient should be evaluated by a healthcare provider for further investigation and management.
- Additional imaging or tests may be required to determine the cause of the opacification in the right lung field.

This report is for informational purposes only and should not be considered a substitute for a professional medical opinion.

Figure 8: **MediX-R1 - Report Generation: Case 2**