Supplementary Material of OraPO: Oracle-educated Reinforcement Learning for Data-efficient and Factual Radiology Report Generation

Zhuoxiao Chen[1,2†], Hongyang Yu[1⋆], Ying Xu[1], Yadan Luo[2], Long Duong[1], Yuan-Fang Li[1⋆]
[1]Oracle Health & AI, [2]The University of Queensland
{ivan.chen, hongyang.yu, ying.x.xu, long.duong, yuanfang.li}@oracle.com
{zhuoxiao.chen, y.luo}@uq.edu.au

This supplementary material includes:

- **Algorithm 1:** A detailed, step-by-step description of the proposed method.
- **Section 1:** Comprehensive implementation details.
- **Section 2:** Additional experimental results and analysis on the MIMIC-CXR dataset.
- **Section 3:** Qualitative studies, including examples of generated reports and assessments on extracted facts.

1. More Implementation Details

Table 1 summaries the hyperparameters of our method and the ranges explored during tuning. We use a small effective batch size of $B = 16$, chosen to fit our compute-efficient 4×A10 setup, and a conservative learning rate of 2.5×10^{-7} to stabilise GRPO updates under high reward variance. The GRPO group size is set to $K = 8$, which we found to provide a good balance between exploration of diverse rollouts and stable gradient estimates. For the ZRR-controlled mixing, we tune the EMA momentum α in $0.4, 0.5, 0.6$ and select $\alpha = 0.5$ so that the zero-reward rate reacts quickly enough to persistent failures, while still smoothing out step-to-step noise. The DPO mixing weights are constrained to a relatively narrow band, with $w_{\min} = 0.05$ and $w_{\max} = 0.15$. In practice, we observed that larger $w_{\max}$ values cause DPO to dominate the update, which slows GRPO exploration and makes the overall behavior resemble supervised fine-tuning rather than reinforcement learning. Conversely, keeping a non-zero $w_{\min}$ ensures that DPO remains available as a gentle corrective signal even when GRPO rewards are mostly informative. Finally, we choose a relatively sharp exponent $\gamma = 2.0$ for mapping ZRR to the mixing weight. This choice makes the weight rise toward $w_{\max}$ only when a prompt experiences many zero-reward groups, so DPO takes over in genuinely hard, unlearned cases, but the weight quickly falls back toward $w_{\min}$ once non-zero rewards appear, allowing GRPO to dominate ordinary updates. We adopt two length-aware variants to handle variable-length reports: DR.GRPO [12] to mitigate length bias in on-policy updates, and LN-DPO [1] to normalise preference margins by sequence length.

Table 1. Hyperparameter settings of the proposed method (selected in underline).

Parameter	Description	Search Range / Setting
B	Effective Batch Size	$\{\underline{16}\}$
LR	Learning Rate	$\{1e-6, \underline{2.5e-7}\}$
K	GRPO sampled group size	$\{4, \underline{8}, 16\}$
α	EMA momentum for zero-reward rate (ZRR)	$\{0.4, \underline{0.5}, 0.6\}$
$w_{\min}$	Minimum DPO mixing weight	$\{0.02, \underline{0.05}, 0.1\}$
$w_{\max}$	Maximum DPO mixing weight	$\{\underline{0.15}, 0.3\}$
γ	Sharpening exponent for mapping ZRR to $w_i^{(t)}$	$\{1.0, \underline{2.0}\}$

Table 2. Experimental results (micro averaging) on the MIMIC-CXR dataset [3].

Algorithm	Venue	Precision, Recall, F1	Train Size
MedRAT [5]	ECCV24	0.285, 0.265, 0.227	223K
MET [16]	CVPR23	0.364, 0.309, 0.311	223K
KiUT [6]	CVPR23	0.371, 0.318, 0.321	369K
DCL [7]	CVPR23	0.471, 0.352, 0.373	223K
RGRG [14]	CVPR23	0.524, 0.474, 0.498	223K
CoFE [8]	ECCV24	0.489, 0.370, 0.405	223K
COMG [4]	WACV24	0.424, 0.291, 0.345	223K
MPO [17]	AAAI25	0.436, 0.376, 0.353	223K
MAN [13]	AAAI24	0.411, 0.398, 0.389	223K
Med-LLM [11]	MM24	0.412, 0.373, 0.395	223K
B-LLM [9]	AAAI24	0.465, 0.482, 0.473	223K
EKAGen [2]	CVPR24	0.517, 0.483, 0.499	223K
MambaXray-L [15]	CVPR25	0.561, 0.460, 0.505	1.27M
MLRG [10]	CVPR25	0.549, 0.468, 0.505	240K
Ours	–	0.342, 0.811, 0.481	1K

2. More experimental results and analysis on MIMIC-CXR

Table 2 complements our main macro-averaged results by reporting micro-averaged Precision, Recall, and F1 on MIMIC-CXR. Macro averaging, used in the main paper, treats each disease label equally and highlights performance

† Work done during an internship at Oracle Health & AI. ⋆ Joint senior authors.

Algorithm 1 The proposed OraPO with FastS Reward for Radiology Report Generation

Input initial policy model π_θ^{init}; dataset $\mathcal{D}$ with studies $(x, y^\star)$ and prompts p; label set L; group size K; hyperparameters $\varepsilon, \lambda_{\text{KL}}, \tau, \alpha, \gamma, w_{\min}, w_{\max}, \beta, \xi$; outer loops I, inner steps M.
Output π_θ

policy model $\pi_\theta \leftarrow \pi_\theta^{\text{init}}$
for iteration $= 1, \ldots, I$ **do**
  reference model $\pi_{\text{ref}} \leftarrow \pi_\theta$
  for step $= 1, \ldots, M$ **do**
    Sample a minibatch $\{(x_i, p_i, y_i^\star, z_i^\star)\}^B \subset \mathcal{D}$
    Update the behaviour policy $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
    Sample K reports $\{\hat{y}_{i,j}\}_{j=1}^K \sim \pi_{\theta_{\text{old}}}(\cdot \mid x_i, p_i)$ for each input radiology image x_i in the batch
    Compute FactS reward: $\{r(x_i, \hat{y}_{i,j})\}_{j=1}^K$ for each sampled report $\hat{y}_{i,j}$;
    for OraPO iterations $= 1, \ldots, \mu$ **do**
      Compute GRPO objective: $L_{\text{GRPO}}(x_i, p_i)$;
      Compute DPO objective: $L_{\text{DPO}}(x_i, p_i)$;
      Compute OraPO objective: $\mathcal{L}_{\text{OraPO}}$ by mixing $L_{\text{DPO}}(x_i, p_i)$ and $L_{\text{GRPO}}(x_i, p_i)$;
      Update the policy model π_θ by minimizing $L_{\text{OraPO}}(x_i, p_i)$.
    end for
  end for
end for

on rare findings. Micro averaging instead weights labels by their frequency, answering a different question: across all positive label instances in the dataset, how many do we miss. This view is especially focusing on common conditions which dominate the label distribution.

Under this micro-averaging metric, OraPO remains highly competitive while using 2–3 orders of magnitude less data. With only 1 K training reports, OraPO attains a micro recall of 0.811 and micro F1 of 0.481. The best baseline recall is EKAGen at 0.483, so OraPO improves micro recall by 67.9% relative to this strong model. Clinically, such high recall means that, across all truly abnormal cases, our method misses far fewer findings, which is critical because false negatives on common diseases are usually more harmful than extra false positives.

Although our method achieves the highest recall, its micro F1 remains close to that of the strongest fully supervised approaches. The best baseline F1 is 0.505 (MambaXray-L and MLRG), whereas OraPO reaches 0.481, only a gap of 0.024. This trade-off is achieved with extreme data efficiency: compared with methods trained on the full 223K MIMIC-CXR split, OraPO uses 99.6% fewer labeled reports, and compared with MambaXray-L's 1.27M training pairs, about 99.9% fewer. Overall, the micro results show that OraPO delivers clinically preferred high recall with competitive F1, while training with only 1K samples.

3. Qualitative Study

In Figure 1, we present three examples that demonstrate how OraPO with FactS reward generates clinically accurate reports while maintaining factual grounding through atomic fact extraction and entailment checking.

Example 1 showcases ideal performance on a challenging case with cardiomegaly and edema. The model produces a clinically coherent narrative that captures the typical imaging pattern of cardiac-related pulmonary edema. Generated facts such as "cardiac silhouette is enlarged" and "diffuse interstitial opacities seen centrally" directly map to the ground-truth labels, with no false positives.

Example 2 highlights OraPO's handling of complex cases with five ground-truth labels. The generated report correctly identifies all target pathologies, including the challenging bilateral nodular pulmonary lesions. While the model predicts pleural effusion absent from the ground-truth labels, the generated facts ("pleural effusions present" with "decreased interval changes") suggest a false positive predicted by the model.

Example 3 illustrates robust multi-pathology detection spanning four conditions: lung opacity, consolidation, pneumonia, and pleural effusion. The model generates anatomically precise descriptions ("left perihilar opacity with patchy distribution") that provide localization detail often absent in baseline systems. All ground-truth labels receive explicit supporting facts, demonstrating the FactS reward's effectiveness in maintaining high recall across co-occurring pathologies. The model correctly attributes findings to specific lung regions, which aids radiologist review.

Across these examples, OraPO consistently produces factually grounded reports where each diagnostic statement traces back to verifiable atomic facts that entail specific labels. This stands in contrast to fluency-optimized base-

lines that generate plausible prose without guaranteed factual alignment. The FactS reward's sentence-level supervision enables the model to maintain clinical accuracy even when handling rare pathologies or multi-label complexity, directly addressing the recall-precision trade-off critical in clinical deployment.

Here's a concise qualitative analysis paragraph you can place right below Fig. 1:

3.1. Qualitative NLG Metric Analysis

Although the ROUGE-L F1 values in Figure 1 are modest, they largely **reflect surface mismatch rather than clinical errors**: OraPO is not trained to imitate the phrasing or template of GT reports. Instead of imitating the GT writing style, our method directly optimizes a fact-level reward. The model first extracts atomic clinical statements, and then checks whether these statements are entailed by the ground-truth labels. This training objective naturally prioritizes clinical correctness and verifiability, not surface-level similarity. In contrast, reference-overlap metrics such as ROUGE mainly measure wording overlap. They reward matched phrasing but do not penalize factual mistakes, missing findings, or even cross-sentence contradictions in long reports. Concretely, our FactScore-based reward treats the report as its own rationale, maps sentences to label-level predictions, and yields dense, interpretable feedback to the policy. In the three qualitative cases, the generated statements are clinically grounded and aligned with the GT labels: Example 1 captures cardiomegaly and edema with supporting facts; Example 2 correctly identifies all target pathologies (with one flagged false-positive effusion); and Example 3 localizes a left perihilar opacity and supports consolidation/pneumonia/effusion—all consistent with the label set despite low lexical overlap with GT wording. This behavior is consistent with our quantitative results. OraPO is deliberately recall-oriented, which is clinically the safer operating point. On CheXpert Plus, it achieves state-of-the-art label-level recall and F1 while using only about 0.1% of the usual training data, and it also attains very high micro-recall. Therefore, low n-gram overlap with the reference should **not be interpreted as low clinical faithfulness**.

Image	GT Reports	GT Labels	Our Generated Reports	Predicted Labels and Supporting Facts	ROUGE-L F1
	Single frontal view of the chest demonstrates indistinct pulmonary vasculature bilaterally, likely representing mild pulmonary edema. Mild cardiomegaly is seen, unchanged from prior exam. No soft tissue or bony abnormalities are seen.	Cardiomegaly, Edema	Ribs: The thoracic rib contour is smooth and complete. There is mild opacity seen in the mid to left hemithorax. Pectoral Girdle: Both clavicles are seen in their normal positions. Diaphragm: The left diaphragm appears increased in curvature without significant deviation. Aortic Arch/Trunk: The aortic arch appears well defined. Cardiomediastinum: There are diffuse interstitial opacities seen centrally on the image, consistent with pulmonary edema. The right border of the heart is poorly defined and the cardiac silhouette is enlarged. The pulmonary vascularity is increased centrally but diffuse through the lung fields and the cardiac silhouette appears to be expanding to the right of the sternum. Pulmonary Vasculature: The pulmonary vasculature appears increased centrally. Lung Opacity: There is diffuse sub-pleural opacity centrally, with extension towards the left hemithorax. Increased pulmonary vasculature seen centrally, consistent with pulmonary edema. No other areas of consolidation seen. This appearance is consistent with pulmonary edema. There are small sub-pleural opacities seen centrally.	- Cardiomegaly: 'Cardiac silhouette is enlarged' and 'Cardiac silhouette appears to be expanding to the right of the sternum' both indicate cardiomegaly. - Edema: 'Diffuse interstitial opacities seen centrally', 'Pulmonary vascularity is increased centrally', 'Pulmonary vascularity is diffuse through the lung fields', and 'Appearance is consistent with pulmonary edema' all indicate edema."	0.102
	Within the right middle lobe there is a large consolidation, consistent with pneumonia, which obscures the right heart border. A small left upper lobe calcified granuloma is present. There is mild calcification of the aortic arch. No pleural effusions. Mild cardiomegaly. No bony abnormalities are appreciated.	Cardiomegaly, Lung Opacity, Lung Lesion, Consolidation, Pneumonia	Cardiomediastinal silhouette: mildly increased in size; cardiac contour is flattened Nodular pulmonary lesions present bilaterally - no growth or interval increase Pleural effusions present - decreased interval changes Pneumonic opacities present, right mid and left lobe, bilateral, but limited in extent Right diaphragmatic elevation	- Cardiomegaly: 'Cardiomediastinal silhouette is mildly increased in size' directly supports presence of cardiomegaly. - Lung Opacity: Facts 'Pneumonic opacities present in right mid lobe', 'Pneumonic opacities present in left lobe', and 'Nodular pulmonary lesions present bilaterally' all describe opacities. - Lung Lesion: 'Nodular pulmonary lesions present bilaterally' directly supports presence of lung lesions. - Consolidation: Facts 'Pneumonic opacities present in right mid lobe', 'Pneumonic opacities present in left lobe', and 'Pneumonic opacities are bilateral' support presence of consolidation. - Pneumonia: Facts 'Pneumonic opacities present in right mid lobe', 'Pneumonic opacities present in left lobe', and 'Pneumonic opacities are bilateral' support presence of pneumonia. - Pleural Effusion: 'Pleural effusions present' directly supports presence of pleural effusion.	0.114
	Initial chest radiograph at 1130 hours demonstrates left lower lobe focal air space opacity which could represent pneumonia. The remainder of the lungs are clear. The cardiomediastinal silhouette is normal. There is blunting of the left costophrenic angle which could represent a small pleural effusion. Follow-up radiograph 1530 hours demonstrates no interval change.	Lung Opacity, Consolidation, Pneumonia, Pleural Effusion	Single ventricular image. Cardiac silhouette is no grossly enlarged (LVED: 55-57 mm, LVED: 55-57 mm) and changes relative to previous films. No evidence of marked cardiomegaly in the presence of mild respiratory straining. Left perihilar opacity with patchy distribution. Small pleural effusion present, mild volume and interval increase noted.Cardiac silhouette interval increase noted.	- Lung Opacity: Fact 'Left perihilar opacity with patchy distribution' describes an opacity, confirming existence. - Consolidation: Fact 'Left perihilar opacity with patchy distribution' describes an opacity consistent with consolidation. - Pneumonia: Fact 'Left perihilar opacity with patchy distribution' describes an opacity that could represent pneumonia. - Pleural Effusion: Facts 'Small pleural effusion present', 'Mild volume increase in pleural effusion', and 'Interval increase in pleural effusion' all confirm pleural effusion.	0.164

Figure 1. X-ray image and its corresponding ground-truth, along with the output of our model generation report on the ChexPert Plus dataset. The mismatch sentence in the reports are highlighted using different colors.

References

[1] Kian Ahrabian, Xihui Lin, Barun Patra, Vishrav Chaudhary, Alon Benhaim, Jay Pujara, and Xia Song. A practical analysis of human alignment with *po. In *Findings of the Association for Computational Linguistics (NAACL)*, pages 8013–8021, 2025. 1

[2] Shenshen Bu, Taiji Li, Yuedong Yang, and Zhiming Dai. Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14194–14204, 2024. 1

[3] Pierre J. 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: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *CoRR*, abs/2405.19538, 2024. 1

[4] Tiancheng Gu, Dongnan Liu, Zhiyuan Li, and Weidong Cai. Complex organ mask guided radiology report generation. In *Proc. Winter Conference on Applications of Computer Vision (WACV)*, pages 7995–8004, 2024. 1

[5] Elad Hirsch, Gefen Dawidowicz, and Ayellet Tal. Medrat: Unpaired medical report generation via auxiliary tasks. In *Proc. European Conference on Computer Vision (ECCV)*, pages 18–35, 2024. 1

[6] Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19809–19818. IEEE, 2023. 1

[7] Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3334–3343, 2023. 1

[8] Mingjie Li, Haokun Lin, Liang Qiu, Xiaodan Liang, Ling Chen, Abdulmotaleb Elsaddik, and Xiaojun Chang. Contrastive learning with counterfactual explanations for radiology report generation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 162–180, 2024. 1

[9] Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. Bootstrapping large language models for radiology report generation. In *Proc. Conference on Artificial Intelligence (AAAI)*, pages 18635–18643, 2024. 1

[10] Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10348–10359, 2025. 1

[11] Rui Liu, Mingjie Li, Shen Zhao, Ling Chen, Xiaojun Chang, and Lina Yao. In-context learning for zero-shot medical report generation. In *Proc. International Conference on Multimedia (MM)*, pages 8721–8730. ACM, 2024. 1

[12] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *CoRR*, abs/2503.20783, 2025. 1

[13] Hongyu Shen, Mingtao Pei, Juncai Liu, and Zhaoxing Tian. Automatic radiology reports generation via memory alignment network. In *Proc. Conference on Artificial Intelligence (AAAI)*, pages 4776–4783. AAAI Press, 2024. 1

[14] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7433–7442, 2023. 1

[15] Xiao Wang, Fuling Wang, Yuehang Li, Qingchuan Ma, Shiao Wang, Bo Jiang, and Jin Tang. Cxpmrg-bench: Pre-training and benchmarking for x-ray medical report generation on chexpert plus dataset. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5123–5133, 2025. 1

[16] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 11558–11567, 2023. 1

[17] Ting Xiao, Lei Shi, Peng Liu, Zhe Wang, and Chenjia Bai. Radiology report generation via multi-objective preference optimization. In *Proc. Conference on Artificial Intelligence (AAAI)*, pages 8664–8672. AAAI Press, 2025. 1