This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

LRX-Bench: Robustness Evaluation of an RLVR-Trained Medical Vision–Language Model under Deployment-Motivated Chest X-ray Degradations

Boniface Chukwuemeka Godwin and Oyindamola Olatunji

Carnegie Mellon University Africa, Kigali, Rwanda
{bgodwin,obolatun}@andrew.cmu.edu

Abstract. Reinforcement learning with verifiable rewards (RLVR) is increasingly used to post-train medical vision–language models (VLMs), but tests typically use clean images. We present LRX-Bench, a controlled stress-testing protocol with five synthetic, SSIM-targeted perturbations across five severity levels. We compare InfiMed-RL-3B, post-trained with reflective supervised fine-tuning (SFT) followed by RLVR, with its Qwen2.5-VL-3B-Instruct base checkpoint on 500 public non-African radiographs (250 TB-positive; 250 normal). Clean accuracies are only slightly above chance (55.8% and 56.8%), but the checkpoints have opposite class biases; these results diagnose failure profiles rather than deployment-ready screening performance. Under low-resolution stress, InfiMed-RL-3B shifts monotonically toward "No TB," reaching 100% negative predictions and 0% TB sensitivity at severity level 4; under dust-like occlusion, its negative-prediction rate reaches 99% at level 5. We call this observed model- and prompt-specific convergence toward B/"No TB" *negative-class collapse*. Because B always denotes "No TB," disease-class, answer-letter, and option-position bias remain competing explanations. The base checkpoint records numerically higher point-estimate accuracy in 23 of 25 perturbed conditions. Without the intermediate SFT-only checkpoint, the comparison characterises the complete InfiMed post-training pipeline rather than isolating a causal effect of RLVR. Similar clean accuracy can therefore conceal substantially different failure profiles under controlled degradation.

Keywords: Medical Vision–Language Models · Reinforcement Learning with Verifiable Rewards · Robustness Evaluation · Chest Radiography · Tuberculosis · Resource-Constrained Imaging

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has become an important post-training approach for improving the reasoning and answer-selection behaviour of language and vision–language models after DeepSeek-R1 [1]. In medical imaging, recent systems such as MedVLM-R1 [3], GMAI-VL-R1 [4], Med-R1

[5], and InfiMed [2] apply reinforcement learning to medical visual question answering and multimodal reasoning. InfiMed-RL-3B combines reflective-pattern-injected supervised fine-tuning (SFT) with RLVR on a Qwen2.5-VL-3B-Instruct backbone and reports gains across seven medical benchmarks. However, these gains are measured primarily on curated benchmark images. Controlled evidence also suggests that medical reinforcement learning (RL) can sharpen an output distribution without necessarily expanding the model's underlying visual support [6], making failure behaviour under image degradation an important open question.

Radiography in resource-constrained health systems can be affected by heterogeneous equipment, poor maintenance, variable acquisition practice, and limited quality-assurance capacity [15, 17]. Such factors can alter spatial resolution, exposure, noise, artefacts, and patient positioning. These conditions are particularly relevant for tuberculosis (TB), for which chest radiography and computer-aided detection are used extensively in high-burden settings [14]. Robustness studies of medical VLMs have begun to evaluate common and modality-specific corruptions [12, 13, 8]; nevertheless, the behaviour of an RLVR-trained medical VLM under a deployment-motivated chest-radiography stress-test suite has not been directly characterised.

We introduce LRX-Bench, a controlled evaluation protocol that adopts Structural Similarity Index Measure (SSIM)-guided severity calibration and applies five deployment-motivated synthetic perturbations to TB chest radiographs. We evaluate InfiMed-RL-3B against its Qwen2.5-VL-3B-Instruct base model while holding the prompt, preprocessing, and decoding procedure fixed. Importantly, InfiMed-RL-3B differs from the base checkpoint through both reflective SFT and RLVR; therefore, this comparison evaluates the complete post-training pipeline and does not isolate RLVR as the sole cause of any observed difference.

The paper makes three contributions. First, it provides a reproducible chest-radiography stress-testing protocol with per-image, SSIM-targeted severity calibration. Second, it reports a case study showing that two checkpoints with similar clean accuracy can exhibit opposite class biases and markedly different degradation trajectories. Third, it documents an observed model- and prompt-specific convergence of InfiMed-RL-3B toward the fixed B/"No TB" response under selected detail-degrading perturbations, while examining alternative explanations and limitations.

2 Related Work

2.1 Reinforcement-Learning Post-Training for Medical VLMs

Medical VLM post-training increasingly combines supervised fine-tuning (SFT) with reinforcement learning. MedVLM-R1 [3] uses reinforcement learning to encourage explicit medical visual reasoning, while GMAI-VL-R1 [4] and Med-R1 [5] investigate RL-based generalisation across medical tasks and modalities. InfiMed [2] first performs reflective-pattern-injected SFT on 188K samples and

then applies RLVR on 36K samples using group relative policy optimisation. It reports an average gain from 57.1% for its SFT checkpoint to 59.2% for its RLVR checkpoint across seven benchmarks. A recent controlled analysis finds that SFT tends to expand task support whereas RL often sharpens the probability mass over already-supported answers [6]. These studies motivate evaluating not only clean accuracy but also whether post-training changes error concentration and failure modes under distribution shift.

2.2 Robustness Evaluation for Medical Foundation Models

Cheng et al. [12] studied medical VLM sensitivity to image artefacts, and Imam et al. [13] introduced corruption-based evaluation across several medical modalities. MedFM-Robust [8] provides 40 perturbations across eight modalities and evaluates VQA, grounding, captioning, and segmentation with five SSIM-guided severity levels. Its base perturbations already include noise, pixelation, brightness changes, and geometric transformations. LRX-Bench does not claim that all the image distortions are new. Instead, it provides a chest X-ray evaluation protocol with settings motivated by deployment-relevant image-quality and workflow constraints, including dust-like obstruction and off-angle imaging. It uses these distortions to compare a publicly available RLVR-trained model with its base model.

2.3 African Chest-Radiography Data Access

African clinical validation requires data collected within African health systems. During the study, we did not obtain access to an African chest-radiography cohort with both TB labels and lesion-level localisation annotations suitable for this evaluation. Recent Ethiopian and Ugandan efforts [18, 19] demonstrate important progress, but their cohorts were unavailable for external validation here. Consequently, LRX-Bench uses public non-African datasets and should be interpreted as a methods and stress-testing study, not as evidence of clinical validity in African populations.

3 LRX-Bench

3.1 Benchmark Composition and Models

The benchmark contains 500 images, balanced between 250 TB-positive and 250 normal examinations. It draws from TBX11K [9] (150 active-TB and 125 normal images from the validation split) and NLM Shenzhen [10] (100 TB-positive and 125 normal images). For Shenzhen, lesion bounding boxes are obtained by fitting rectangles to the released pixel-level masks of TB-consistent abnormalities [11]. TBX11K stores one row per annotated box; we deduplicate by image filename before stratified sampling, reducing 248 active-TB annotation rows to 164 distinct images. All images are resized to 512×512 before model input.

Ethics and data governance. This secondary analysis used publicly released chest radiographs and involved no new participant recruitment or data collection. Shenzhen was de-identified by its providers; its NIH use and public release were exempted from IRB review (OHRP No. 5357). The TBX11K publication and official release do not report an ethics-committee or informed-consent determination, so we make no further claim about the underlying collection.

We evaluate **InfiMed-RL-3B** [2] and **Qwen2.5-VL-3B-Instruct** [7] zero-shot. The checkpoints share the Qwen2.5-VL-3B architecture, but InfiMed-RL-3B has undergone reflective SFT followed by RLVR. Thus, observed differences are associated with the complete InfiMed post-training pipeline and are not attributed to RLVR alone.

3.2 Perturbation Suite and SSIM Calibration

LRX-Bench includes five synthetic perturbations motivated by image-quality and workflow constraints in resource-limited radiography (Table 1). Following MedFM-Robust [8], binary search selects per-image parameters targeting five SSIM bands. From levels 1 to 5, the bands are 0.90–0.98, 0.80–0.89, 0.70–0.79, 0.60–0.69, and 0.50–0.59. Calibrated parameters are cached and reused across model evaluations.

LRX-Bench uses controlled, deployment-motivated perturbations to represent broad categories of degradation relevant to resource-limited radiography. SSIM-based calibration provides a common image-similarity scale across images, enabling comparison under targeted bands, subject to the reported low-resolution level-5 shortfall. The perturbations are not device-specific physical simulations; they form a reproducible stress test for failure behaviours that clean-image evaluation may obscure.

3.3 Evaluation Protocol and Metrics

The VQA prompt is: "Is there evidence of tuberculosis in this chest X-ray? (A) Yes (B) No. Answer with only A or B." The grounding prompt requests a bounding box in $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$ normalised to $[0, 1]$, or `null` when no lesion is present. Grounding is requested for all 500 images to characterise output behaviour, including box emission on normal scans.

We report overall and class-conditional VQA accuracy, the B-answer rate (fraction mapped to "No TB"), and clean-to-perturbed accuracy change. Clean-image confidence intervals use 10,000 bootstrap resamples. Perturbed comparisons are descriptive point estimates because paired uncertainty was not estimated. For grounding, we record IoU@0.5 and normal-image box emission, but do not interpret them as localisation performance under the current prompt and parser.

Table 1. LRX-Bench perturbations and their deployment motivation. The perturbations provide controlled, SSIM-targeted approximations of relevant image-quality degradations.

Type	Mechanism and motivation
Low resolution	Downsample–upsample cycle used to attenuate fine spatial detail and stress sensitivity to reduced effective resolution.
Dust-like occlusion	Random opaque circular patches used as a simplified proxy for detector-surface contamination and occluding artefacts [16].
Exposure variation	Bidirectional gamma correction around γ=1 used to stress sensitivity to broad exposure and tone variation.
Off-axis perspective	Perspective warp used as a geometric stress test motivated by positioning variability and constrained radiography workflows [17].
Portable-noise proxy	Combined Gaussian and salt-and-pepper noise used to stress sensitivity to mixed high-frequency corruption.

4 Results

4.1 Clean-Image Performance and Prediction Bias

Table 2 summarizes clean and perturbed performance. Both checkpoints achieve modest clean accuracy (55.8% and 56.8%), only a few percentage points above the 50% chance level on this balanced binary task, with overlapping bootstrap confidence intervals. These results do not support deployment-ready TB screening performance; the analysis instead compares failure profiles under controlled degradation. However, the same aggregate accuracy conceals opposite class biases. InfiMed-RL-3B predicts "No TB" for 65% of clean images, yielding 70.8% accuracy on normal images and 40.8% on TB-positive images. The base checkpoint predicts "Yes TB" for 65% of images, yielding 72.0% accuracy on TB-positive images and 41.6% on normal images (Fig. 2). The InfiMed post-training pipeline is therefore associated with an inversion of the clean-image class bias relative to the base checkpoint, without a corresponding improvement in overall accuracy on this task.

InfiMed reports average accuracies of 49.2% for Qwen2.5-VL-3B and 59.2% for InfiMed-RL-3B across seven medical benchmarks [2]. On LRX-Bench's binary TB task, the clean gap is only one percentage point (55.8% versus 56.8%). This difference in task behaviour reinforces the need for task- and deployment-specific evaluation.

Table 2. Zero-shot LRX-Bench point estimates (500 images; five severity levels per perturbation). Δ: clean accuracy minus mean perturbed accuracy; negative values indicate higher mean accuracy after perturbation. Worst: largest decrease.

Model	Clean Acc	95% CI	Δ	Worst
InfiMed-RL-3B	0.558	[0.514, 0.602]	0.011	0.080
Qwen2.5-VL-3B	0.568	[0.526, 0.612]	−0.011	0.066

4.2 Prediction Behaviour under Perturbation

The base model records numerically higher point-estimate VQA accuracy than InfiMed-RL-3B in 23 of 25 perturbed conditions (mean perturbed accuracy: 0.579 versus 0.547; Fig. 1). These comparisons are descriptive; paired uncertainty was not estimated.

Under **low-resolution** stress, InfiMed-RL-3B's B-answer rate rises monotonically from 70% to 86%, 96%, 100%, and 100% across severity levels 1–5. At level 4, it predicts "No TB" for every image, producing 0% sensitivity to TB-positive cases. The reported calibration shortfall applies only to the most severe low-resolution target, level 5; the headline result is already reached at level 4. Because the prompt fixes B as "No TB," the experiment cannot distinguish disease-negative preference from answer-letter or option-position bias. Under **dust-like occlusion**, the B-answer rate reaches 99% at level 5. We use *negative-class collapse* for this observed model- and prompt-specific convergence toward B/"No TB," not for a general property of post-trained VLMs. The base checkpoint's low-resolution B-answer rates are 38%, 57%, 55%, 83%, and 92%, showing a later, non-monotonic shift. The other perturbations do not produce the same monotonic trajectory. The strongest observed shift is therefore associated with removal or occlusion of fine image detail.

4.3 Grounding Output Analysis

Under the present prompt/parser configuration, grounding is *uninterpretable as a measure of localisation ability.* IoU@0.5 is 0.000 for both checkpoints, but this cannot separate absent spatial capability from coordinate non-compliance or parser mismatch. InfiMed-RL-3B emits boxes on 31.2% of normal images (78 of 250), while the base emits none; these are output-behaviour statistics, not localisation-performance measurements.

5 Discussion

Interpreting the observed collapse. InfiMed-RL-3B combines reflective SFT with RLVR, and controlled work suggests that RL can sharpen an output distribution when task support already exists [6]. The monotonic low-resolution trajectory is compatible with post-training sharpening of the checkpoint's clean-image B/"No TB" preference as visual evidence degrades. It does not identify

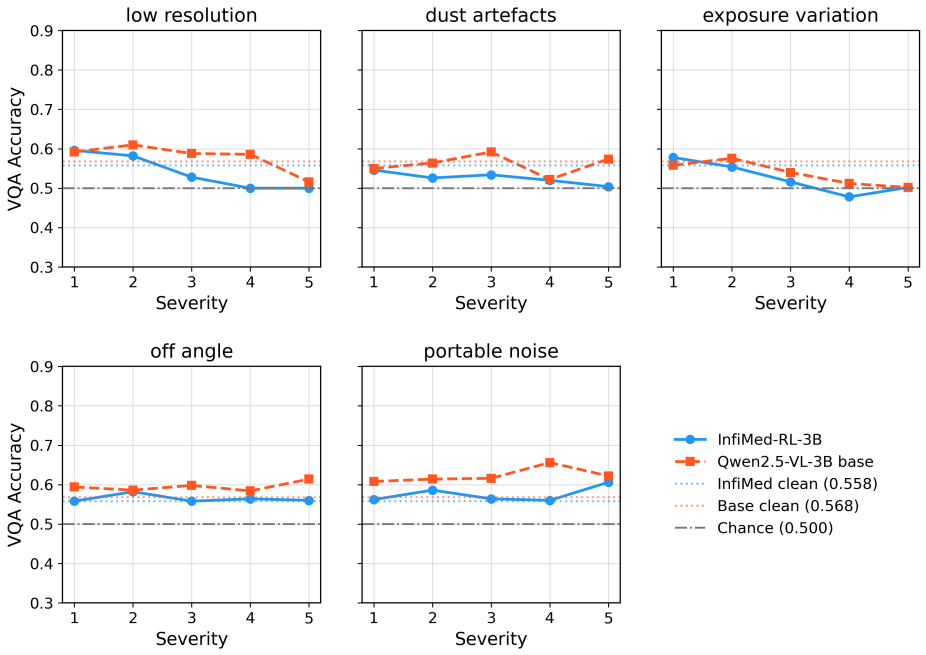

Fig. 1. VQA accuracy versus severity level for all five LRX-Bench perturbation types. Model curves are identified in the legend. Dotted lines indicate clean-image accuracy per model; the dashed line marks chance (0.50). The largest point-estimate divergence occurs under low-resolution and dust-like occlusion stress. No inferential comparison is made for the perturbed conditions.

RLVR as causal because InfiMed-SFT-3B is absent, and the fixed B/"No TB" mapping leaves answer-letter and option-position bias unresolved. We therefore use *negative-class collapse* only for the observed pattern in this model–prompt pipeline.

The perturbation-specific pattern is also compatible with increased dependence on fine visual detail: low-resolution and dust-like occlusion remove or cover local texture, whereas the tested photometric and geometric transformations preserve more of the image's local structure. This interpretation remains provisional because the perturbations are synthetic proxies and no frequency- or region-based analysis is reported.

Implications for medical VLM evaluation. Nearly identical clean accuracy can hide clinically asymmetric error profiles: one checkpoint favours sensitivity, while the other favours specificity and can lose all TB sensitivity under selected degradations. Because both clean accuracies are only slightly above chance, this is diagnostic evidence about failure behaviour, not clinical or deployment readiness. Medical-VLM robustness reports should therefore include class-conditional performance and output-bias diagnostics. Future grounding work

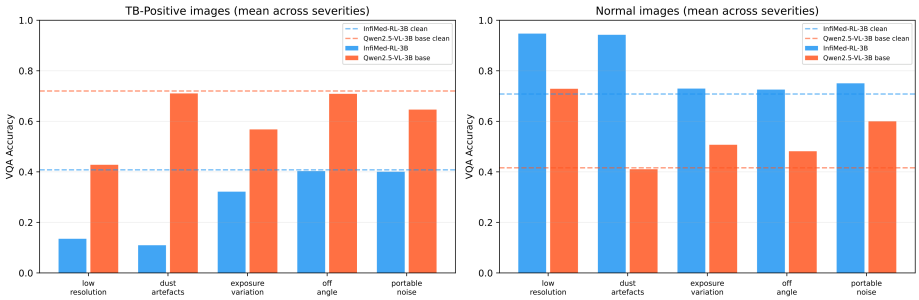

Fig. 2. Per-class accuracy averaged across all severity levels. Left panel: TB-positive images; right panel: normal images. Within each perturbation pair, the left bar is InfiMed-RL-3B and the right bar is Qwen2.5-VL-3B, preserving interpretability in grayscale. The checkpoints exhibit opposite class biases.

should verify parse validity and coordinate conventions before reporting spatial performance.

6 Limitations

This study evaluates one post-trained checkpoint and one base checkpoint. Because InfiMed-RL-3B includes reflective SFT and RLVR, the design does not isolate RLVR's causal contribution; the SFT-only checkpoint is the key missing control. The perturbations are synthetic rather than measured device-specific corruption models, and equal SSIM does not guarantee equal loss of clinical information.

Per-image calibration missed the severity-level-5 low-resolution SSIM target for approximately 30% of images, so level-5 findings require caution. This shortfall does not affect the headline level-4 observation, where InfiMed-RL-3B already reaches 100% "No TB" predictions and 0% TB sensitivity. Statistical uncertainty is reported only for clean accuracy; perturbed differences are not significance tests. The fixed option order does not separate disease-class bias from answer-letter or position bias. Source images are not from African facilities, so the results cannot establish clinical performance or fairness in African populations. Shenzhen rectangles fitted to pixel masks may be imprecise for diffuse disease. Grounding remains uninterpretable because coordinate or parser mismatch cannot be separated from spatial failure. The models are not certified medical devices, and this robustness study is not clinical validation.

7 Conclusion

LRX-Bench stress-tests a post-trained medical VLM under synthetic degradations of chest radiographs. In this balanced 500-image evaluation, InfiMed-RL-3B and its base checkpoint have near-chance clean accuracy but opposite class

biases. Under low-resolution stress, InfiMed-RL-3B converges monotonically toward B/"No TB" and reaches 0% TB sensitivity; under dust-like occlusion, its B-answer rate reaches 99% at level 5. We call this observed model- and prompt-specific pattern negative-class collapse, not a general property of RLVR-trained VLMs. The evidence associates the complete InfiMed post-training pipeline with a different failure profile but does not isolate RLVR or establish deployment readiness. The single most informative next step is evaluation of InfiMed-SFT-3B, followed by prompt-order tests, paired uncertainty, additional post-trained models, and mitigation studies. Evaluation on African clinical cohorts also remains necessary.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Guo, D., Yang, D., Zhang, H., et al.: DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning. Nature **645**, 633–638 (2025). https://doi.org/10.1038/s41586-025-09422-z
2. Liu, Z., Hou, Z., Zhu, G., Sang, Z., Xie, C., Yang, H.: InfiMed: Low-Resource Medical MLLMs with Advancing Understanding and Reasoning. arXiv:2505.23867 (2025)
3. Pan, J., Liu, C., Wu, J., Liu, F., Zhu, J., Li, H.B., Chen, C., Ouyang, C., Rueckert, D.: MedVLM-R1: Incentivizing Medical Reasoning Capability of Vision–Language Models (VLMs) via Reinforcement Learning. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. LNCS, vol. 15966, pp. 337–347. Springer, Cham (2025). https://doi.org/10.1007/978-3-032-04981-0_32
4. Su, Y., Li, T., Liu, J., et al.: GMAI-VL-R1: Harnessing Reinforcement Learning for Multimodal Medical Reasoning. arXiv:2504.01886 (2025)
5. Lai, Y., Zhong, J., Li, M., Zhao, S., Yang, X.: Med-R1: Reinforcement Learning for Generalizable Medical Reasoning in Vision–Language Models. arXiv:2503.13939 (2025)
6. Jeddi, A., Shaban, K., Baghbanzadeh, N., et al.: When Does RL Help Medical VLMs? Disentangling Vision, SFT, and RL Gains. arXiv:2603.01301 (2026)
7. Bai, S., Chen, K., Liu, X., et al.: Qwen2.5-VL Technical Report. arXiv:2502.13923 (2025)
8. Cui, X., Huang, T., Wang, Y., Hu, L., Yin, L.: MedFM-Robust: Benchmarking Robustness of Medical Foundation Models. arXiv:2605.19027 (2026)
9. Liu, Y., Wu, Y.H., Ban, Y., Wang, H., Cheng, M.M.: Rethinking Computer-Aided Tuberculosis Diagnosis. In: Proc. IEEE/CVF CVPR, pp. 2646–2655 (2020)
10. Jaeger, S., Candemir, S., Antani, S., Wang, Y.X.J., Lu, P.X., Thoma, G.: Two Public Chest X-Ray Datasets for Computer-Aided Screening of Pulmonary Diseases. Quant. Imaging Med. Surg. **4**(6), 475–477 (2014). https://doi.org/10.3978/j.issn.2223-4292.2014.11.20
11. Yang, F., Lu, P.X., Deng, M., WÃ¡ng, Y.X.J., Rajaraman, S., Xue, Z., Folio, L.R., Antani, S.K., Jaeger, S.: Annotations of Lung Abnormalities in the Shenzhen Chest X-ray Dataset for Computer-Aided Screening of Pulmonary Diseases. Data **7**(7), 95 (2022). https://doi.org/10.3390/data7070095

12. Cheng, Z., Ong, A.Y., Wagner, S.K., et al.: Understanding the Robustness of Vision–Language Models to Medical Image Artefacts. npj Digital Medicine **8**(1), 727 (2025). https://doi.org/10.1038/s41746-025-02108-w
13. Imam, R., Marew, R., Yaqub, M.: On the Robustness of Medical Vision–Language Models: Are they Truly Generalizable? In: Medical Image Understanding and Analysis. MIUA 2025. LNCS, vol. 15916, pp. 233–256. Springer, Cham (2026). https://doi.org/10.1007/978-3-031-98688-8_17
14. World Health Organization: Global Tuberculosis Report 2025. Geneva (2025). https://www.who.int/teams/global-programme-on-tuberculosis-and-lung-health/tb-reports/global-tuberculosis-report-2025
15. International Atomic Energy Agency: Worldwide Implementation of Digital Imaging in Radiology. Human Health Series No. 28, IAEA, Vienna (2015). https://www.iaea.org/publications/10613/worldwide-implementation-of-digital-imaging-in-radiology
16. Walz-Flannigan, A.I., Brossoit, K.J., Magnuson, D.J., Schueler, B.A.: Pictorial Review of Digital Radiography Artifacts. RadioGraphics **38**(3), 833–846 (2018). https://doi.org/10.1148/rg.2018170038
17. Kawooya, M.G., Kisembo, H.N., Remedios, D., et al.: An Africa Point of View on Quality and Safety in Imaging. Insights into Imaging **13**, 58 (2022). https://doi.org/10.1186/s13244-022-01203-w
18. Gebrekidan, M., Hailu, S.S., Abera, M.T., Asmare, M.H., Negussie, M.A.: Multicenter Chest Radiographic Findings from Ethiopia with Radiologist-Generated Annotations. Research Square (2026). https://doi.org/10.21203/rs.3.rs-9417593/v1
19. Musinguzi, D., Murindanyi, S., Katumba, A.: PaliGemma-CXR: A Multi-task Multimodal Model for TB Chest X-ray Interpretation. arXiv:2503.00171 (2025)