# Faithful Reasoning, Not More Reasoning: Chain-of-Thought Ablation for Tuberculosis Findings in Indonesian Chest X-Rays

Luthfi Nur Hanfan[1], Giovanni Montana[2], Cindy Adelia Setiawan[3], Reyhan Eddy Yunus[4], and Vanya Valindria[1*]

[1] Data Science, Faculty of Information Technology, Monash University, Indonesia
[*] vanya.valindria@monash.edu
[2] Department of Statistics, University of Warwick, UK
[3] Independent Researcher
[4] Department of Radiology, Dr. Cipto Mangunkusumo National General Hospital, Faculty of Medicine, Universitas Indonesia, Indonesia

**Abstract.** Tuberculosis (TB) screening from chest X-rays is a setting where automated reading could ease severe shortages of radiologists, and vision-language models (VLMs) now attempt it while showing their reasoning step by step through chain-of-thought (CoT) prompting. Recent radiology models report accuracy gains from such reasoning, but mostly after extra training and measured as overall report quality, so whether prompting alone helps a model pick out findings, which reasoning step matters, and whether this holds outside English remain open. We study TB on chest X-rays with radiologist reports in Indonesian. Using one strong general-purpose VLM, we compare direct prompting against a base CoT and six arms, five changing one of its reasoning steps and one combining two, scoring six TB findings for accuracy (F1) and for false positives against the report, with paired significance tests. No CoT arm improves F1 over direct prompting. Adding a single verification step, in which the model re-checks each finding it reported, is the one clear gain: it significantly lowers false positives at no accuracy cost, and the findings it does report are well grounded in the model's own earlier observations. This self-grounding explains why some prompts over-report more than others.

**Keywords:** Chest X-ray · Tuberculosis · Vision-language models · Chain-of-thought · Hallucination

## 1 Introduction

Indonesia carries the world's second-highest tuberculosis (TB) burden and the largest national gap between estimated and reported cases [33, 28]. The gap is driven partly by a shortage of readers: chest X-ray (CXR) is the frontline screening test [8, 29], yet Indonesia has about 1.2 radiologists per 100,000 people against a benchmark near 7.6 [36]. With the World Health Organization now

permitting computer-aided detection in place of a human reader for TB CXR screening [32], automated reading at scale is a realistic target, provided it works on reports written in Bahasa Indonesia.

Such a tool must do more than return a label: a clinician who acts on or overrides an automated read needs to see which signs it claims and on what evidence, the more so because CXR reading is itself reader-dependent [8, 22]. Vision-language models (VLMs) promise this kind of explainable reading, and recent radiology systems add explicit chain-of-thought (CoT) reasoning to make the explanation auditable [7, 20, 15, 35, 21]. Whether that reasoning genuinely grounds the reported findings or merely narrates around them is unsettled, and for screening it is a safety question: a model that reads confidently while fabricating signs is more dangerous than a terse one. We evaluate five open VLMs for Indonesian TB findings extraction: the general-domain Qwen3.5 and four medically fine-tuned models, MedGemma, CheXagent-2, VILA-M3, and CXR-LLaVA-v2. Qwen3.5 leads the medical models: on a broader 14-class TB slice under the same structured prompt, its macro F1 is 0.297, against 0.221 for MedGemma, 0.170 for CheXagent-2, 0.153 for VILA-M3, and 0.118 for CXR-LLaVA-v2. That slice uses a different cohort and protocol, so it is not comparable with the results reported below. Qwen3.5 and MedGemma are also the only two to follow the structured Indonesian prompt at all, while CheXagent-2, VILA-M3, and CXR-LLaVA-v2 revert to English free-text, native captioning, or a collapsed template. We therefore focus on the sharpest contrast, the generalist Qwen3.5 against the clinically tuned MedGemma. The medical model is not merely less accurate: under a CoT instruction it destabilizes into long English reasoning traces and language-inconsistent output, so we run the single-factor ablation on the stable generalist.

On that base we ask whether CoT helps over direct prompting and, if so, which step and why. We run a single-factor ablation: a shared base CoT and six arms from published radiology-reasoning methods, five changing exactly one step and one combining two. We score every condition with established metrics (F1 and a false-positive over-affirmation rate) and with the metric we propose, a self-grounding rate (SGR) measuring how often a reported finding is backed by the model's own earlier observations, all under paired significance tests.

We position our hallucination rate and the SGR against two lines of prior work. The first scores fabrication against a reference: object-hallucination benchmarks such as POPE [13] and CHAIR [27], and radiology tools that re-ground generated measurements [10]. Against these we report a reference-free over-affirmation rate over the whole report. The second scores reasoning faithfulness, but only at extra cost: by perturbing and re-running the model (Turpin et al. [31], Lanham et al. [12]), resampling it (SelfCheckGPT [16]), reading its internal signals (Xie et al. [34]), or requiring reference labels or white-box access (ReXTrust [9], ChestX-Reasoner [7], Moll et al. [19]). The SGR requires none of this machinery: in a single pass, with no extra runs, no ground truth, and no second model, it checks whether each finding the model reports is supported by

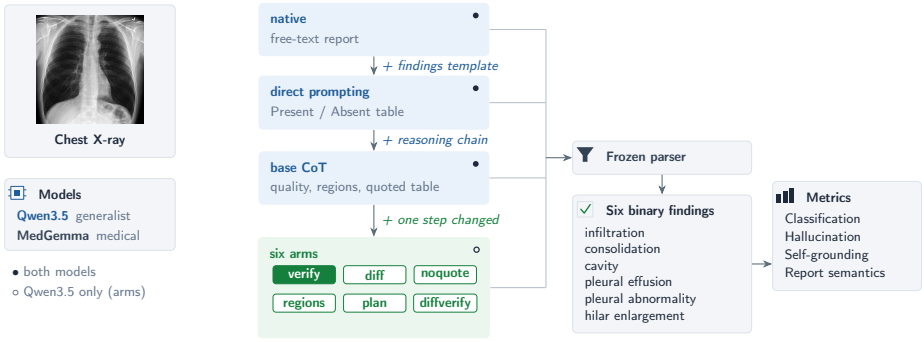


**Fig. 1.** Evaluation pipeline. A chest radiograph with an Indonesian findings prompt enters three base conditions (native, direct, base CoT) plus six arms; the base conditions run on both models, the arms on Qwen3.5 only (filled vs. open dots). One frozen parser scores all conditions identically.

a phrase from its own earlier description. We interpret it alongside hallucination and omission, as MedHEval [2] pairs a hallucination measure with recall.

Our contributions are: (1) an evaluation of open VLMs for Indonesian TB findings extraction, in which a general-domain model leads the medical ones, which collapse on the language task; (2) a single-factor CoT ablation showing that a single verification step cuts hallucination at no F1 cost, while no arm improves F1 over direct prompting; and (3) the self-grounding rate (SGR), a new ground-truth-free metric that explains these over-affirmation differences rather than merely ranking them.

## 2 Method

### 2.1 Data and Ground Truth

We evaluate on anonymized chest radiographs (DICOM) from Rumah Sakit Cipto Mangunkusumo (RSCM), Jakarta, Indonesia, curated and provided through the Big Data Center (BDC), IMERI, Faculty of Medicine, Universitas Indonesia. Each study is a single frontal radiograph paired with a free-text radiologist report in Indonesian. We restrict the label space to the six tuberculosis-relevant findings active in this cohort: infiltration, consolidation, cavity, pleural effusion, pleural abnormality, and hilar enlargement. Reference labels come from an automatic report labeler [4] binarized at probability 0.5; because it operates on English, each Indonesian report is machine-translated to English first, while model predictions are parsed directly from the Indonesian output. The labeler, not manual re-annotation, is the reference standard. We evaluate all 600 studies, one per patient, from a single cohort and year, running every prompting condition on the identical image set so all comparisons are paired by study.

### 2.2 Models and Reporting Setup

We study two open-weight VLMs: *Qwen3.5* (9B) [26], a general-domain model, and *MedGemma* (4B) [30], fine-tuned on clinical data. Both are multimodal, so the contrast is domain specialization, not modality. Each model receives the radiograph and an instruction to report the six findings in Indonesian, with deterministic greedy decoding. We compare three base prompting conditions, *native* (free text), *direct prompting* (a Present/Absent table), and *base CoT* (a reasoning chain before the table), plus six *arms*, each the base CoT with one step changed, except `diffverify`, which combines two (Fig. 1). The three base conditions run on both models, but the six arms run on Qwen3.5 alone, because under a CoT instruction MedGemma destabilizes and its findings table parses with incomplete coverage, so an arm's effect on it would be confounded with base-model instability rather than the prompt step. Qwen3.5 is the stronger, better-calibrated base, and keeping to one base fits our compute scope.

### 2.3 Chain-of-Thought Prompt Arms

The base CoT elicits three steps before the answer: a brief image-quality judgment; a free-text description of each anatomical region; and the Present/Absent table, in which each Present call must quote the phrase from the description step that supports it. Each arm (Fig. 1) modifies exactly one of these steps, so any behavior change is attributable to that single factor; `diffverify` alone combines two: `verify` [15] re-checks each Present call and drops the unsupported; `diff` [25] adds a location and one or two differential diagnoses to the description; `noquote` [20] drops the supporting-quote requirement; `regions` [20, 35] replaces the description with a twelve-region checklist; `plan` [7] prepends a planning preamble assuming normal unless proven; and `diffverify` [15, 25] combines `diff` and `verify`. Exact prompts and the arm-to-source mapping are in the code repository: `https://github.com/hanfan-nur/CoT_Ablation_Indonesian_TB`.

### 2.4 Findings Extraction

From each response, a single deterministic parser, frozen across all conditions, extracts the six binary findings from the model's own structured table (Fig. 2), so differences between arms reflect the prompt, not the reader. A finding the model does not mark Present is treated as Absent (default-Absent). Coverage is complete for Qwen3.5, but for MedGemma the parser needed hardening to read its table formatting, and a fraction of its cells revert to English. The native condition produces no table, so it enters only the semantic comparison (Section 2.5). Model and reference findings pass through different extractors, so extraction errors are not shared between prediction and ground truth.

### 2.5 Evaluation Metrics

We report four axes. **Classification:** per-finding F1 and macro F1 (the mean of the six per-finding F1 scores), plus precision and recall. Because the six-finding

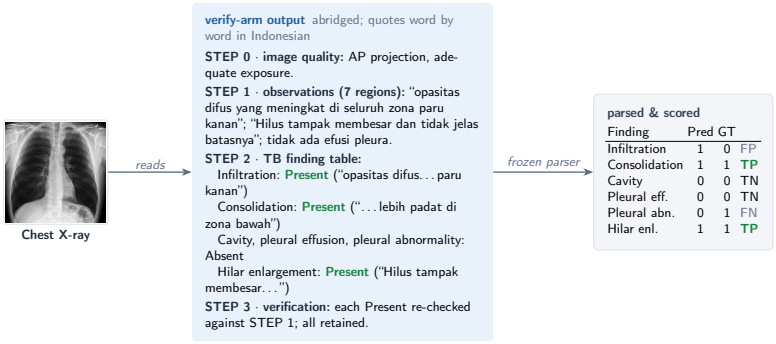


**Fig. 2.** Worked example (real output). A chest radiograph (left) is read by the `verify` arm in four steps, and a frozen parser maps its output to six binary predictions scored against the report ground truth. STEP 3, which re-checks each Present call against the model's own STEP 1 observations, is the basis of the SGR.

grid is dominated by absent labels, which inflates plain accuracy, we report balanced accuracy (the mean of recall and specificity, so an all-Absent predictor scores 0.5) and also report accuracy and the Matthews correlation coefficient [17]. **Hallucination and omission:** a false-positive over-affirmation rate $\mathrm{FPR}_h = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$, POPE-style [13]; the instance false-discovery rate $\mathrm{FDR}_h = \mathrm{FP}/(\mathrm{TP} + \mathrm{FP})$, CHAIR-style [27]; the omission rate $\mathrm{FNR} = \mathrm{FN}/(\mathrm{TP} + \mathrm{FN})$, so that silence is not rewarded [2, 11]; an over-call ratio $(\mathrm{TP} + \mathrm{FP})/(\mathrm{TP} + \mathrm{FN})$, predicted-positive calls over reference-positive findings (1 at calibration, above 1 over-reporting, below 1 under-reporting); and a study-level rate, the fraction of images with at least one false positive. **Semantics:** BLEU [23, 24], ROUGE [14], and BERTScore [37, 3] between the generated and the radiologist report, computed on a narrative scope (the free-text description prose) so the comparison is not confounded by the extra table and verification text that some arms emit. **Self-grounding:** for each Present call we test whether the phrase the model quotes appears in its own earlier description step. The SGR is that grounded fraction, an operational, reference-free measure of one facet of reasoning faithfulness [31, 12, 16, 34, 9, 7, 19].

### 2.6 Statistical Significance

All comparisons are paired by image; for each arm we test its change against the base CoT. Per finding we use McNemar's test on the discordant pairs (an exact binomial for small counts, a continuity-corrected $\chi^2$ otherwise) [18, 5], with a second McNemar on reference-negative items to isolate hallucination, and we control the false discovery rate across the arm-by-finding grid with Benjamini-Hochberg [1]. For each macro delta we attach a 95% confidence interval from 1,000 image-level bootstrap resamples [6]. We keep an arm only when it lowers the over-affirmation rate with a confidence interval excluding zero while its F1 interval still includes zero.

**Table 1.** Qwen3.5 (generalist) vs. MedGemma (medical) on the six TB findings, 600 images, default-Absent macro scores. Best per column in bold; lower is better for $FPR_h$. MedGemma's lower $FPR_h$ reflects severe under-calling (low recall), not better calibration. Bal. acc.: balanced accuracy; $FPR_h$: false-positive over-affirmation rate; Prec., Rec.: precision, recall.

| Model | Prompt | F1 | Prec. | Rec. | Bal. acc. | $FPR_h$ |
|---|---|---|---|---|---|---|
| Qwen3.5 (generalist) | direct | **0.336** | **0.419** | **0.372** | **0.607** | 0.158 |
| Qwen3.5 | CoT | 0.309 | 0.394 | 0.351 | 0.583 | 0.184 |
| MedGemma (medical) | direct | 0.178 | 0.405 | 0.143 | 0.545 | **0.053** |
| MedGemma | CoT | 0.200 | 0.298 | 0.235 | 0.538 | 0.159 |

**Table 2.** Qwen3.5 CoT arm ablation, 600 images, default-Absent macro scores. † marks the recommended arm (`verify`); bold marks our proposed metric (SGR). Against base CoT, `verify` significantly lowers hallucination ($FPR_h$) at no significant F1 cost, while `diff` and `diffverify` significantly raise it; significance tests are in the text, and `plan`'s low $FPR_h$ reflects under-calling (over-call 0.56), not better grounding. Over-call is (TP+FP)/(TP+FN), predicted positives over reference positives; $FDR_h$ is the instance false-discovery rate FP/(TP + FP); SGR is the grounded fraction of Present calls.

| Condition | F1 | $FPR_h$ | $FDR_h$ | Over-call | **SGR** |
|---|---|---|---|---|---|
| Direct prompting | 0.336 | 0.158 | 0.581 | 1.24 | n/a |
| Base CoT | 0.309 | 0.184 | 0.606 | 1.35 | 0.82 |
| `verify`† | 0.322 | 0.148 | 0.557 | 1.17 | 0.86 |
| `diff` | 0.325 | 0.253 | 0.653 | 1.73 | 0.77 |
| `noquote` | 0.320 | 0.176 | 0.575 | 1.26 | n/a |
| `regions` | 0.289 | 0.141 | 0.630 | 1.03 | 0.78 |
| `plan` | 0.236 | 0.068 | 0.510 | 0.56 | 0.88 |
| `diffverify` | 0.321 | 0.226 | 0.650 | 1.61 | 0.78 |

## 3 Results and Discussion

**Baseline: Generalist versus Medical.** The generalist outperforms the medically fine-tuned model on every accuracy column of Table 1, scoring highest under plain direct prompting, and the gap persists under chain-of-thought. MedGemma's weakness is concentrated in recall, and it emits English throughout where Qwen3.5 reports in Indonesian. Chain-of-thought pushes both to call more findings, with opposite effects: for the near-calibrated generalist the extra calls are mostly false positives; for the under-calling medical model they raise recall, lifting macro F1 but at nearly triple the false-positive rate. Neither is a clean gain, and only the generalist is both accurate and calibrated. MedGemma also destabilizes under chain-of-thought, opening every output with an English reasoning block (no Qwen3.5 output does) at roughly triple the median length.

**Chain-of-Thought Does Not Improve F1.** No chain-of-thought condition significantly improves macro F1 over direct prompting (Table 2, macro over the six findings on the common 600 images): the six arms cluster near the base CoT,

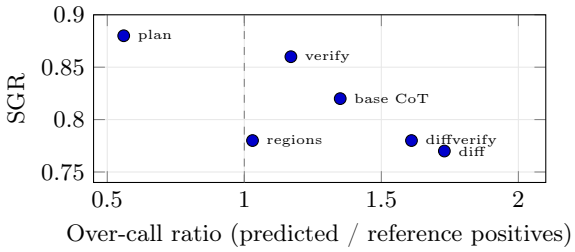


**Fig. 3.** SGR versus over-call across the Qwen3.5 CoT conditions (600 images): grounded arms sit at low over-call, elaborating arms over-call more; the dashed line marks calibration. `plan` sits high only because it barely calls findings Present (over-call 0.56). `noquote` is absent, having no quotes to ground.

none significantly above it, and the only significant F1 move is `plan`'s, downward ($\Delta$F1 $= -0.073$).

**Verification Cuts Hallucination at No F1 Cost.** Against the base CoT, `verify` significantly lowers the false-positive rate ($\Delta\text{FPR}_h = -0.036$) while its F1 interval includes zero, making it the one clear keep, and it posts the highest Matthews correlation coefficient of any condition (0.212); `regions` also lowers $\text{FPR}_h$ significantly but gains no F1. In the opposite direction, `diff` and `diffverify` significantly raise it, over-calling most. The `plan` arm looks safest on $\text{FPR}_h$ but is the omission trap: it misses most findings (FNR 0.81), so omission must be read beside hallucination at a screening venue. The `noquote` arm changes nothing measurable.

**Faithful Reasoning Explains the Difference.** The SGR explains the over-affirmation pattern rather than merely ranking beside it (Fig. 3). Among the arms that still answer, `verify` is the most self-grounded and near-calibrated, so it affirms fewer unsupported findings; `diff` is the least self-grounded and the worst over-caller, as elaborating the reasoning induces rationalization. Because the SGR scores only the calls a model makes, not how many it makes, `plan` reaches a high value purely by staying silent (the omission trap), so the SGR must be read beside the omission rate, not alone. The `noquote` arm removes the quoting requirement by design and so has no SGR.

**Report Semantics.** On a narrative scope that isolates the free-text observation prose, the CoT reports are as similar to the radiologist report as direct prompting: sentence BLEU is 0.69 for direct and, `regions` aside, 0.73 to 0.84 across the CoT conditions, with ROUGE-1 no lower. The exception is `regions`, whose verbose twelve-region sweep is least report-like (BLEU 0.49). BERTScore sits at its noise floor and does not discriminate. Reasoning therefore does not cost report quality on this scope; the full per-metric table is in the code repository.

**Limitations.** Several limitations bound these claims. The ablation varies one base model, so the effects hold for a strong, well-calibrated generalist and may differ on other backbones. Both the reference labels and the predictions come from automatic labelers not yet validated against a human gold standard, so a false positive is hallucination relative to the report, its translation, and the labeler, not proven fabrication; a human-labeled gold subset is the next rigor gate. The SGR uses exact-substring matching, a strict lower bound that penalizes paraphrase and so understates grounding for the more verbose arms. Finally, the over-affirmation rate is report-relative: a finding the report omits counts against the model even when present on the image. We state these so our central claim, that more faithful (better-grounded) reasoning reduces fabrication, is read as a within-model, paired comparison, not an absolute accuracy claim.

## 4 Conclusion

On tuberculosis findings extraction, chain-of-thought prompting does not improve F1 over direct prompting; it changes the error profile instead. A single self-verification step, our `verify` arm, significantly reduces over-affirmation at no F1 cost, while requesting differential diagnoses (the `diff` arm) increases it; our new ground-truth-free self-grounding rate (SGR) explains both, since an arm cuts fabrication when it grounds its conclusions in its own observations. Faithful reasoning, not more reasoning, is the safer prompt; whether the SGR transfers across backbones is the natural next test.

**Acknowledgments.** The chest X-ray scans were collected retrospectively under approval from the Health Research Ethics Committee, Faculty of Medicine, Universitas Indonesia and RSCM (approval KET-68/UN2.F1/ETIK/PPM.00.02/2026, 16 January 2026). This study was funded by the Monash Warwick Alliance Research Activation Fund, Round one 2025, under the project "Enhancing AI-Based TB Detection: Automated Radiology Report Generation from Chest X-ray in Indonesia".

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

## References

1. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) **57**(1), 289–300 (1995). https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
2. Chang, A., et al.: MedHEval: Benchmarking hallucinations and mitigation strategies in medical large vision-language models. arXiv:2503.02157 (2025)
3. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th ACL. pp. 8440–8451. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.747

4. Dicente Cid, Y., et al.: Development and validation of open-source deep neural networks for comprehensive chest x-ray reading: a retrospective, multicentre study. The Lancet Digital Health **6**(1), e44–e57 (2024). https://doi.org/10.1016/S2589-7500(23)00218-2
5. Edwards, A.L.: Note on the correction for continuity in testing the significance of the difference between correlated proportions. Psychometrika **13**(3), 185–187 (1948). https://doi.org/10.1007/BF02289261
6. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall/CRC (1993)
7. Fan, Z., Liang, C., Wu, C., Zhang, Y., Wang, Y., Xie, W.: ChestX-Reasoner: Advancing radiology foundation models with reasoning through step-by-step verification. arXiv:2504.20930 (2025)
8. Hansun, S., Argha, A., Liaw, S.T., Celler, B., Marks, G.: Machine and deep learning for tuberculosis detection on chest X-rays: Systematic literature review. Journal of Medical Internet Research **25**, e43154 (2023). https://doi.org/10.2196/43154
9. Hardy, R., Kim, S.E., Ro, D.H., Rajpurkar, P.: ReXTrust: A model for fine-grained hallucination detection in AI-generated radiology reports. arXiv:2412.15264 (2024)
10. Heiman, A., et al.: FactCheXcker: Mitigating measurement hallucinations in chest X-ray report generation models. In: CVPR (2025)
11. Kim, Y., et al.: Medical hallucinations in foundation models and their impact on healthcare. arXiv:2503.05777 (2025)
12. Lanham, T., et al.: Measuring faithfulness in chain-of-thought reasoning. arXiv:2307.13702 (2023)
13. Li, Y., et al.: Evaluating object hallucination in large vision-language models. In: Proceedings of EMNLP. pp. 292–305. Association for Computational Linguistics (2023). https://doi.org/10.18653/v1/2023.emnlp-main.20
14. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics (2004)
15. Luo, Y., He, W., Cui, Z.X., Liang, D.: Teaching AI stepwise diagnostic reasoning with report-guided chain-of-thought learning (DiagCoT). arXiv:2509.06409 (2025)
16. Manakul, P., Liusie, A., Gales, M.J.: SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In: Proceedings of EMNLP. pp. 9004–9017. Association for Computational Linguistics (2023). https://doi.org/10.18653/v1/2023.emnlp-main.557
17. Matthews, B.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) – Protein Structure **405**(2), 442–451 (1975). https://doi.org/10.1016/0005-2795(75)90109-9
18. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947). https://doi.org/10.1007/BF02295996
19. Moll, J., et al.: Evaluating reasoning faithfulness in medical vision-language models using multimodal perturbations. In: Proceedings of Machine Learning for Health (ML4H). Proceedings of Machine Learning Research, vol. 297 (2025)
20. Myronenko, A., et al.: Reasoning visual language model for chest X-ray analysis (NV-Reason-CXR-3B). arXiv:2510.23968 (2025)
21. Ng, C., Sun, L., Tang, S.: X-Ray-CoT: Interpretable chest X-ray diagnosis with VLMs via chain-of-thought reasoning. arXiv:2508.12455 (2025)
22. Nguyen, H.Q., et al.: VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. Scientific Data **9**, 429 (2022). https://doi.org/10.1038/s41597-022-01498-w

23. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th ACL. pp. 311–318. Association for Computational Linguistics (2002). https://doi.org/10.3115/1073083.1073135
24. Post, M.: A call for clarity in reporting BLEU scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers (WMT). pp. 186–191. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/W18-6319
25. Prakash, K.D., Kim, K., Han, Y.: Enhancing clinical reasoning in medical vision-language model through structured prompts. medRxiv (2025). https://doi.org/10.1101/2025.07.31.25331901
26. Qwen Team: Qwen3.5: Towards native multimodal agents (2026), `https://qwen.ai/blog`
27. Rohrbach, A., et al.: Object hallucination in image captioning. In: Proceedings of EMNLP. pp. 4035–4045. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/D18-1437
28. Saktiawati, A.M.I., Probandari, A.: Tuberculosis in Indonesia: challenges and future directions. The Lancet Respiratory Medicine **13**, 669–671 (2025). https://doi.org/10.1016/S2213-2600(25)00168-7
29. Santosh, K., Allu, S., Rajaraman, S., Antani, S.: Advances in deep learning for tuberculosis screening using chest X-rays: The last 5 years review. Journal of Medical Systems **46**, 82 (2022). https://doi.org/10.1007/s10916-022-01870-8
30. Sellergren, A., et al.: MedGemma 1.5 technical report. arXiv:2604.05081 (2026)
31. Turpin, M., Michael, J., Perez, E., Bowman, S.R.: Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In: NIPS. vol. 36, pp. 74952–74965 (2023)
32. WHO: WHO consolidated guidelines on tuberculosis. module 2: Screening. WHO, Geneva (2021)
33. World Health Organization: Global tuberculosis report 2025. WHO, Geneva (2025)
34. Xie, Z., Guo, J., Yu, T., Li, S.: Calibrating reasoning in language models with internal consistency. In: NIPS (2024)
35. Yao, Y., et al.: Thought graph traversal for test-time scaling in chest X-ray VLLMs. Pattern Recognition **179**, 113639 (2026). https://doi.org/10.1016/j.patcog.2026.113639
36. Yunus, R.: Radiology loading and coverage hours in Indonesia. Korean Journal of Radiology **25**(7), 597–599 (2024). https://doi.org/10.3348/kjr.2024.0267
37. Zhang, T., et al.: BERTScore: Evaluating text generation with BERT. In: International Conference on Learning Representations (ICLR) (2020)