

# R.A.I.S.E - A Novel Framework for Evaluating Foundational AI Models in Medical Deployment: Moving Beyond Traditional Metrics to Real-World Deployability

Jacobus AE Adendorff<sup>1</sup>[0009-0002-4790-3672], Laing Lourens<sup>2</sup>[0009-0007-6824-3868], Rhena Delport<sup>3</sup>[0000-0002-6274-0923], Vukosi Marivate<sup>1,4,5</sup>[0000-0002-6731-6267], and Judy W Gichoya<sup>6</sup>[0000-0002-1097-316X]

<sup>1</sup> University of Pretoria

<sup>2</sup> Council for Scientific and Industrial Research NGEI, South Africa

<sup>3</sup> Department of Family Medicine, School of Medicine, University of Pretoria

<sup>4</sup> Data Science for Social Impact

<sup>5</sup> Lelapa AI

<sup>6</sup> Emory University School of Medicine, Atlanta , GA, USA

**Abstract.** The shift from “narrow” traditional deep learning models to more generalist foundation models represents a paradigm shift for AI in medicine with the emergence of unimodal and multimodal systems such as MedGemma, Biomedclip, DINO models, and MedImageInsight. While these generalist models promise broad capabilities, they demand large datasets and high computational resources for training, and carry risks such as hallucinations, which can be hazardous in clinical use. In medicine, whether a model can be securely incorporated into actual clinical workflows is more important than whether it passes standardized tests. Current assessment techniques for foundation models are frequently based on multiple choice questions and do not account for real-world deployment scenarios. At a two-day datathon (16-17 July 2025), we explored deploying MedGemma for chest X-ray reporting in South Africa. We proposed a gradual, radiologist-guided integration focused on controlled, automatable tasks rather than full diagnostic use. Our three-pronged evaluation framework creates a uniform readiness score and allows for continuous real-world monitoring by combining tailored deployment paths and hierarchical decision making with Go/No-Go thresholds.

**Keywords:** Radiology · Implementation Framework · Healthcare · Artificial Intelligence · Foundation Model

## 1 Introduction

The transition from traditional deep learning models to foundational models represents a paradigm shift in artificial intelligence [1], particularly in medical applications with the development of several unimodal and multimodal base models, including MedGemma, Biomedclip, DINO (Distillation with NO labels)

group of models, and MedImageInsight [2, 15]. This shift from semi-supervised to self-supervised learning approaches promises to eliminate the complexity of single task “narrow” models that predict a specific clinical finding, towards generalist models capable of a wide variety of tasks including report generation, report summarization, knowledge retrieval, clinical bot development, and improved interpretability among other tasks [9]. Additionally, these generalist models facilitate efficient training of downstream tasks, which can often be accomplished with limited datasets and modest computational resources [2]. Its this latter ability that promises to leapfrog AI development and deployment for many under-resourced clinical settings in sub-Saharan Africa, including South Africa where there are insufficient datasets and compute resources available, in addition to other infrastructure-related challenges [12].

The emergent abilities of generalist models enable them to be adapted for a task they were not trained for originally by learning from a few examples or by directed guidance using prompting approaches [9]. This characteristic is critical for medical AI due to the need for continuous learning after human feedback, and the need for model improvement for long-tailed tasks - those under-represented or unseen during training. However, the transition from narrow to generalist models presents challenges, as foundational models demand vast datasets and substantial computational resources for their initial training - capabilities typically limited to large corporations. Due to their generative characteristics, these models are susceptible to hallucinations, increasing the risk in medicine for potential harm [15].

Current models, despite their generalist nature, are still unable to generalize to new domains when the task is not represented in the training dataset. For example, knee osteoarthritis prediction performs better on deep learning (DL) models than foundation models due to the lack of knee x-rays in the training datasets of most models. Critically, existing threshold-based approaches for determining the ultimate performance of DL models do not translate to the non-deterministic nature of foundation models. Simply stated, a DL model designed for the task of predicting pneumonia presence or absence often uses a threshold of 0.5 to assess performance probability - a threshold that can be modified to align with the intended sensitivity or specificity of the model. The non-deterministic nature of foundation models makes this impossible, and instead developers have resorted to benchmarking performance using standardised question banks such as USMLE [16] or AfriMedQA [13].

The fundamental question facing clinicians is not whether a model can pass standardized tests like USMLE exams, but whether it can be safely and effectively deployed in specific clinical contexts. Current evaluation frameworks that rely on multiple-choice questions (MCQs) fail to capture real-world deployment scenarios and human interaction patterns, important for successful AI deployment [8]. Additionally, foundation models are difficult to regulate with several governance frameworks in development. For example, the recent development of the European Union AI Act [7] underscores the global movement towards risk-based AI regulations. It places responsibility for safety, transparency, and

local validation on both the developers and deployers - particularly in healthcare where AI systems are classified as high-risk [6]. This shift highlights the need for practical frameworks that support contextual validation and progressive, clinical-driven deployment.

To address these gaps, a multidisciplinary evaluation initiative - structured as a datathon - was conducted, leading to the development of the Radiologist AI Score and Evaluation (RAISE) framework. RAISE addresses three critical gaps in foundational model evaluation: (1) cost-effective assessment without extensive annotation, (2) personalised deployment strategies, and (3) hierarchical decision-making process to enable safe, incremental clinical integration.

## 2 Methodology

The datathon - a structured, collaborative event aimed at rapid, interdisciplinary problem-solving - was modelled after similar events organized by the MIT Critical Data team but adapted for South African context. The methodology emphasized participatory, context-aware evaluation rather than purely quantitative benchmarking. This required creating an environment where technical experts and domain specialists could work together in real time to identify relevant use cases, interrogate model outputs, and iteratively refine their problem framing. The datathon format was selected as it allowed concentrated, collaborative engagement over a short period, enabling participants to explore the intersection of AI capabilities, clinical relevance, and local healthcare priorities.

**Event Design** A two-day datathon was convened on July 16th - 17th 2025, co-organized by the Healthcare AI Innovation and Translation Lab (HITI) at Emory University and the African Institute of Data Science and Artificial Intelligence at the University of Pretoria [10]. The goal of the datathon format was to facilitate rapid mapping of the problem space in collaboration with multidisciplinary experts. Approximately 70 participants were recruited from University of Witwatersrand and University of Pretoria, including representatives from two affiliated hospitals and members of the Data Science Institute’s network. Participants represented a wide range of disciplines, including radiology, paediatrics, critical care, computer science, epidemiology, and public health. The diversity ensured that both technical and clinical perspectives informed the evaluation process. We provided pre-datathon seminars focusing on AI in medicine and common pitfalls of AI [3]. The chosen theme was framed as a single guiding question: “Can we deploy MedGemma for chest X-ray reporting in South Africa?” The evaluation focused on common radiological conditions prevalent in the region, such as tuberculosis (TB) [11], to ensure that use cases reflected realistic local clinical priorities.

To establish a shared understanding between different experts, the datathon began with an overview of RexVQA - a multiple-choice question (MCQ) dataset that has been used for benchmarking different radiology foundation models [14]

- followed by demonstrations of MedGemma using zero-shot prompting, few-shot prompting, in context learning and subsequent fine-tuning. Participants were then randomly assigned to small, multidisciplinary groups to ensure representation across domains. The groups worked independently, receiving regular check-ins from their mentors, culminating in a final presentation of their created projects.

Across multiple model architectures and parameter settings, several foundation models - including MedGemma - performed poorly on basic radiology tasks when prompted with locally relevant cases.

**Framework Development Process** Based on findings of the early evaluation demonstrating multiple pitfalls, a decision was made to move away from attempting to validate whether a foundational model could accurately perform full autonomous diagnosis. Such evaluations carry high clinical risk for the radiologist, making them more hesitant to use the AI system. Additionally, high rate of inaccuracies can erode trust - potentially leading to AI being rejected outright.

Instead, the focus shifted to a more pragmatic goal: helping radiologists navigate an environment of rapidly evolving models, frequent new releases, and non-standard accuracy benchmarks, while building confidence in AI assistance through controlled evaluation.

This led to three guiding principles:

1. **Local validations** - Users should be able to evaluate and compare different models based on tasks relevant to their specific environment.
2. **Customised assistance** - Users should be able to define the emphasis and scope of AI assistance to complement available resources.
3. **Future expansion** - Users should be able to apply the same evaluation criteria to evaluate newer models as they become available.

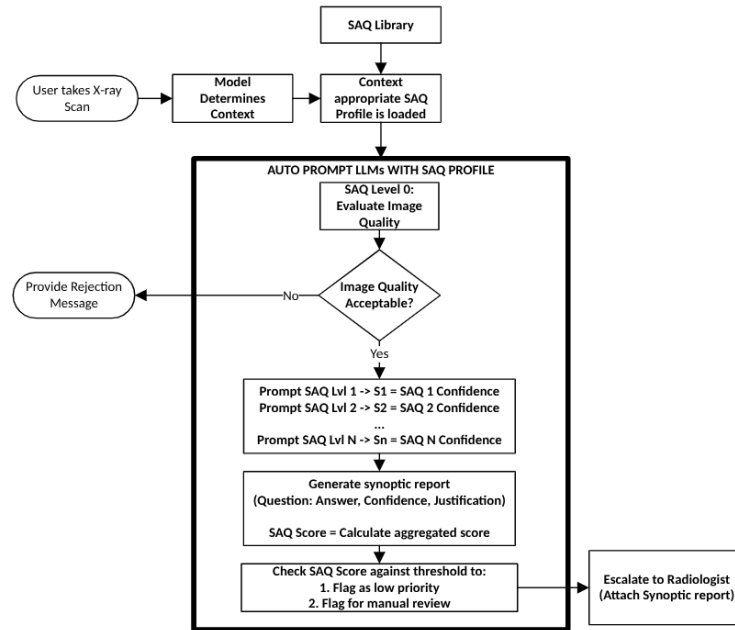
Under this framework, the model is explicitly positioned as an assistant rather than a replacement. The scope of its role is defined and controlled by the radiologist, enabling AI adoption at a pace that they are comfortable with. This paper presents the RAISE framework as a proposal for a controlled, clinically relevant evaluation process. The example in appendix A demonstrates how the approach could be applied in practice, but is not intended to present a statistically comprehensive evaluation.

### 3 Framework

The Radiologist AI Score and Evaluation (RAISE) framework enables radiologists to evaluate and incrementally integrate AI models into their workflows. At its core are context-specific Single-Answer Questions (SAQs), which allows large language models (LLM)-based systems to provide targeted, interpretable support within clearly defined boundaries. These SAQs are organised into levels

of increasing interpretive complexity (ranging from simple detection increasingly advancing to diagnostic inference) and can be combined into profiles that reflect the need of different settings (e.g., clinical trial screening or routine site surveys)

The process entails: prompting a model with progressively complex SAQ levels, compiling the responses into a structured synoptic report for the radiologist, and calculating a weighted, aggregated score for guiding scan prioritisation based on defined thresholds.



**Fig. 1. Workflow with integrated SAQ processing.** The model should first establish the context (which will determine the SAQs used). It then determines whether the image is acceptable. Rejected images are sent back for rescanning. Accepted images are processed through a series of automatic prompting. The weighted confidence scores are aggregated and compared against a threshold for further processing.

The following subsections (3.1 to 3.5) describes the core components of the RAISE framework as illustrated in figure 1.

### 3.1 SAQs: Single-Answer-Questions

Single-Answer-Questions (SAQs) are simple prompts designed to constrain the output of the model. SAQs are categorised into hierarchical levels, while SAQ use cases can be pre-defined using profiles.

**Prompts** Each SAQ is crafted to return three elements:

1. Binary answer - a direct Yes or No response. (The only exception being the very first question where it needs to identify the context of the x-ray).
2. Confidence score - the model's self-reported confidence in its answer, expressed as a percentage (e.g., 92%). These scores are provisional, as self-reported confidence is often miscalibrated and should ideally be validated or adjusted through calibration or uncertainty quantification.
3. Brief justification - a short explanation to support the answer given.

This format offers several practical advantages: It allows model comparisons by standardising the output, enables performance tracking for individual tasks, and constrains the model's scope for automation and other actions. SAQs cover various categories, from quality checks (e.g., "Is the exposure appropriate?") to clinical assessments ("Are the opacities diffuse?"). They are created, selected, and sequenced by the radiologist, ensuring that the implementation is relevant to the local setting and aligned with workflow priorities. Beyond clinical complexity, SAQs can be tailored for targeted evaluations for specific patient subgroups, making the evaluation more relevant to the context where it will be deployed.

**Levels** SAQs are organised into levels of increasing complexity and clinical risk. This tiered structure enables controlled, incremental adoption of AI assistance, with each level building on the validation of the SAQ level before it. An overview of the SAQ levels is displayed in table 1.

Local validation at each level indicates the model's accuracy for that category of task. More cautious radiologists may choose to limit their SAQs to lower levels - expanding to higher levels as confidence grows. Regular validations on updated models can indicate when it is safe to introduce more complex questions.

**Profiles** While SAQ levels define the complexity and risk of individual questions, SAQ profiles determine which combination of levels and questions should be applied to a given clinical context. A profile is essentially a preconfigured set of SAQs custom built for a specific use case.

Profiles can be defined by:

- Image Context - e.g., Chest x-ray vs. pelvis x-ray.
- Operational context - e.g., routine hospital imaging vs. clinical trial imaging.
- Clinical setting - e.g., high-volume TB screening vs. post-operative follow-up.
- Risk tolerance - e.g., Only up to level 1 for cautious deployments.

This allows radiologists to adapt RAISE to different workflows without re-designing the question set or enforcing a fixed, one-size-fits-all automated workflow.

**Table 1.** Structured levels of SAQ questions.

Level	Name	Description and Examples
0	Context Establishment	The only question not constrained to a yes-or-no answer. Its purpose is to determine the appropriate SAQ set for the case at hand. For example, identifying whether the scan is a chest x-ray or pelvis x-ray will determine which prompts follow.
1	Quality Assessment	Questions focus on whether the scan is adequate for use. If the model’s confidence score exceeds a defined threshold indicating inadequate quality, the scan can be rejected early, preventing unnecessary review by the radiologist. Examples include: <ul style="list-style-type: none"><li>– Is the exposure appropriate?</li><li>– Is the patient positioned upright?</li></ul>
2	Low Impact Clinical Observations	Observations that can guide workflow but do not directly trigger diagnostic decisions. Examples include: <ul style="list-style-type: none"><li>– Is a pleural effusion present?</li><li>– Are the opacities diffuse?</li></ul>
3	Moderate Impact (requires clinical reasoning)	Questions that require nuanced interpretation and have a greater potential impact on the patient. Examples include: <ul style="list-style-type: none"><li>– Is there radiographic evidence of active tuberculosis?</li><li>– What is the severity of the pathology?</li></ul>
4	Diagnostic Suggestions	Higher risk that requires subjective judgement. Examples include: <ul style="list-style-type: none"><li>– Do the radiographic findings suggest pulmonary tuberculosis as the most probable diagnosis?</li><li>– Given the image alone, is TB more likely than alternative causes (e.g., bacterial pneumonia, lung cancer)?</li></ul>

### 3.2 Score Calculation

The RAISE framework aggregates SAQ responses into a weighted score combining model confidence and clinical importance. This score guides prioritisation or screens out low-risk cases in resource-limited settings, helping radiologists focus on the most critical cases.

**Formula** For a given scan, the total score is calculated as:

$$\text{Total Score} = \Sigma(\text{confidence}_i \times \text{weight}_i)$$

for all  $i \in S$ , where:

- $S$  is the set of all active SAQs in the active profile
- $\text{confidence}_i$  is the model’s confidence score for SAQ  $i$  (0-100%)
- $\text{weight}_i$  is the importance assigned by the radiologist to SAQ  $i$  (0-1 scale)

Confidence scores are expressed as percentages but converted to decimals when multiplied by the weight. An example of this calculation can be seen in table 2.

**Table 2.** SAQ example calculation - The self-reported confidence expressed by the model is multiplied by the weight (importance) assigned by the radiologist to output a score. This score is aggregated over all the levels and can be used for decision making (e.g., screening out or prioritising patients)

SAQ	Level	Confidence	Weight	Score
Is the exposure appropriate?	1	98	0.20	18.4
Are the opacities diffuse?	2	85	0.25	21.25
Is there radiographic evidence of TB?	3	78	0.45	35.1
<b>Total Score</b>				74.75

**Thresholds and Actions** The total score is compared against predefined thresholds that determine the next step in workflow:

- $\geq 85$  : The scan is flagged as high priority for immediate review.
- 70-84 : The scan is routed for standard review.
- $< 70$  : The scan may safely be screened out or marked as low-priority.

Thresholds are configurable and can be adjusted according to risk tolerance, available resources, or other conditions. By assigning higher weights to more clinically relevant questions, radiologists ensure that the threshold and subsequent flagging aligns with their workflow.

### 3.3 Model Evaluation

Using the RAISE scoring mechanism, radiologists can locally validate and compare models using identical SAQ profiles, enabling more informed, task-specific evaluations.

**Comparative assessment** The binary answers, confidence scores, and justifications are recorded and weighted according to importance. The resulting total scores allow for side-by-side comparison between models to highlight where models perform well or where their accuracies are more unstable. This format is more explicit and relevant to practical use than abstract benchmarks.

Table 3 presents an example comparison between two models (Model A and Model B) evaluated across two subgroups of the population. For each SAQ, average scores are shown for normal (N), abnormal (A), and extreme (E) cases, with a score calculated per level to indicate the model’s performance at each SAQ complexity level.

In this context:

- Normal (N) denotes studies without detectable pathology
- Abnormal (A) denotes studies showing pathology within typical presentation parameters
- Extreme (E) denotes studies with rare, severe, or atypical presentations requiring higher interpretive skill.

**Table 3. Model Evaluation:** Radiologists define an SAQ series relevant to their use and compare the performance of different models. A score for each level is calculated to allow radiologists to see the areas one model outperforms another.

SAQ	Category	Question	Subgroup 1						Subgroup 2					
			Model A			Model B			Model A			Model B		
			N	A	E	N	A	E	N	A	E	N	A	E
1	Quality	Image Alignment	97.4	95.1	92.8	96.6	94.7	92.1	98.2	96	93.1	97.3	95.5	92.7
1	Quality	Exposure	96.3	94.6	91.5	95.7	93.9	90.8	97.1	95	92	96.2	94.2	91.4
1	Quality	Blurring	98	96.4	93.7	97.6	95.8	93.2	98.6	96.8	94.3	98	96.1	93.6
1		<b>Level Score</b>	97.2	95.4	92.7	96.6	94.8	92	98	95.9	93.1	97.2	95.3	92.6
2		Consolidation	93.6	90.4	87.2	92.4	89.9	86.5	94.2	91.1	88	92.9	90.3	87.4

**Ongoing Monitoring** RAISE supports repeated evaluations, serving as a form of instrument validation. This enables radiologists to: (1) track performance changes after model updates or retraining; (2) compare the current production model against new candidate models before replacement; and (3) monitor for performance drift, especially in environments with changing patient demographics or disease prevalence.

### 3.4 Model Implementation

Once a model has been evaluated and its performance validated for a specific SAQ level, it can be introduced into the clinical workflow in a controlled and

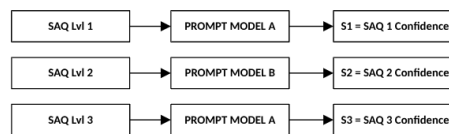
incremented manner. The RAISE framework ensures that the scope of model assistance is explicitly defined, with the radiologist retaining full control over which levels are active at any given time.

**Workflow Integration** During scan processing, the model’s SAQ responses are compiled into a structured synoptic report that accompanies the image for radiologist review. (*See Appendix A for an example*) This report contains: (1) binary answers to each SAQ in the profile, (2) the model’s confidence score for each answer and (3) a brief justification for the answer given. Radiologists can then access the rationale behind each answer and decide whether to accept, ignore, or override it. This approach ensures human oversight while still allowing the model to reduce the administrative work and cognitive load.

**Threshold-Based Automation** In settings with high case volumes or resource constraints, RAISE can be configured to automatically route scans based on the aggregated score.

These routing decisions are strictly governed by thresholds defined by the radiologist, ensuring that automation never exceeds its intended use. As algorithm performance improves and models become better at answering the SAQs accurately, radiologists can add more questions to the library. This lets them unload incrementally more work onto the models, freeing themselves for work that requires more expertise.

**Hybrid Model Configuration** In scenarios where multiple models are available, RAISE supports hybrid prompting - allowing specific SAQs to be assigned to the model best suited to answer that particular question. This concept is illustrated in figure 2.



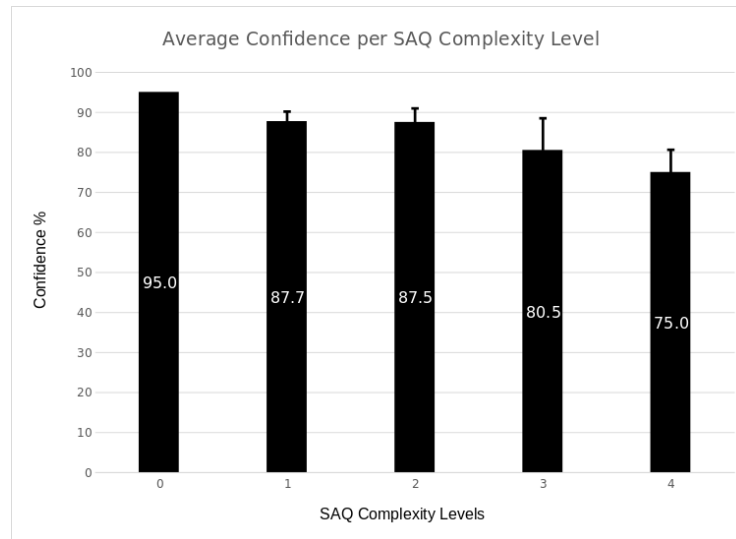
**Fig. 2. Hybrid Prompting:**Radiologists can choose the best performing model for a SAQ level.

### 3.5 Workflow Example Summary

To illustrate the RAISE process in practice, Appendix A shows the output from a foundation model (Claude) when we evaluated an acute pulmonary TB case from PhysioNet’s MIMIC-CXR database [5] using the SAQ framework. We selected

Claude as our demonstration model because it is explicitly listed among the large language models permitted under PhysioNet’s terms of use. Levels 0-4 were applied sequentially, progressing from context, image quality to probable diagnosis. Figure 3 shows the trend of model confidence over each level. The workflow results for each level are summarised as follows:

- Level 0-1: High confidence (95 - 87.6%) for low-risk questions like context, positioning, exposure and blurring checks.
- Level 2: Still above 85% on average confidence for absence of pleural effusion and opacity diffuse, reflecting strong model performance on basic abnormality detection.
- Level 3: The confidence dropped to an average of 80.5% (lowest is 75%) for feature-specific findings (upper lobe opacity, cavitation, lymphadenopathy and severity of pathology). This indicates increased complexity and subjectivity.
- Level 4: Lowest confidence of 74% average for probable TB diagnosis - demonstrating more uncertainty in high-risk tasks.



**Fig. 3.** Model Confidence across SAQ levels - The model’s confidence drops with higher levels containing more complex questions.

## 4 Discussion

We introduce RAISE, a three-pronged evaluation framework designed to bridge the gap between model capabilities and deployment readiness. The framework

provides: (1) structured SAQ-driven assessment that constrains model output to binary responses with justification, thereby reducing the scope for hallucinations, (2) support for personalised deployment profiles, and (3) hierarchical decision-making with multiple Go/No-Go thresholds. These features are aggregated into a single score, similar to threshold levels of deterministic models, which simplifies deployment while enabling continuous real-world monitoring through evaluation dashboards.

This approach moves beyond traditional accuracy metrics to encompass factors such as clinical relevance, safety margins, and interpretability requirements. Recognising that clinical needs vary across institutions, specialities, and practitioners, the framework incorporates personalisation mechanisms that allow for customised thresholds, evaluation criteria, and adaptive learning pathways that respect regulatory constraints. Finally, the hierarchical decision-making enables tiered evaluation: low-level questions handle study identification and basic checks, while only complex cases escalate to higher interpretive levels. For example, an initial question might ask “Is this a Chest X-Ray?” — if the model fails here, processing halts and the case routes to the standard workflow, reducing the risk of bias from downstream misclassification.

## 5 Limitations

This paper presents RAISE as a conceptual framework with a preliminary proof-of-concept. As such, we acknowledge several limitations that also define important avenues for future research:

1. The illustrative example in this paper used a limited set of SAQs, a small case set, and a single foundation model (Claude). To establish the statistical robustness of the RAISE score, future work must involve a comprehensive validation study with a large, diverse dataset (including well-defined Normal, Abnormal, and Extreme cases), multiple competing foundation models (e.g., MedGemma, BiomedGPT), and a more extensive, clinically-validated SAQ library.
2. A core assumption of the current framework is that a model’s self-reported confidence is a reliable proxy for its accuracy. However, model confidence is not always well-calibrated and can vary significantly across models and tasks. Future iterations of RAISE should investigate methods for calibrating these confidence scores or incorporating independent uncertainty quantification techniques to provide a more robust basis for the aggregated score.
3. The strength of RAISE lies in its context-specific SAQ profiles, but their creation requires significant upfront investment of time and clinical expertise from radiologists. Future research could explore semi-automated methods for generating candidate SAQs from clinical guidelines or existing report templates to lower the barrier to adoption.

Despite these limitations, we believe the RAISE framework provides a valuable and pragmatic roadmap for navigating the complex challenge of clinical AI deployment.

## 6 Conclusion

The introduction of foundational models offers transformative potential for AI-assisted healthcare, especially in resource-constrained settings. However, their susceptibility to hallucinations and opaque (black-box) decision-making creates real risks for clinical adoption. Simplified accuracy metrics lack context-specific validation, leaving radiologists without a clear roadmap for integration.

The RAISE framework addresses these challenges through its novel SAQ-driven scoring system, which quantifies model performance across progressive, context-aware prompts and translates these into a weighted, interpretable score. This enables radiologists to compare, validate, and incrementally integrate foundational models into their workflows based on local evidence of reliability, rather than abstract benchmarks. By keeping control in the hands of radiologists and supporting local validation, RAISE aligns with emerging regulations such as the European Union AI Act [7].

As foundational models continue to evolve, frameworks like RAISE can help ensure their adoption remains safe, practical and context-appropriate. RAISE not only provides this mechanism but also establishes a reproducible blueprint for safe, practical, and regulation-ready adoption of AI in radiology.

**Acknowledgments.** The authors gratefully acknowledge the support of the ABSA Chair of Data Science for facilitating this research. DSFSI is supported by the UK International Development and the International Development Research Centre, Ottawa, Canada as part of the AI for Development: Responsible AI, Empowering People Program (AI4D). DSFSI is thankful for gifts from NVIDIA, Google.org, OpenAI and Meta which enable our research. Judy Gichoya receives funding from NIH, grant 1R25OD039834-01. The datathon was supported by the Carnegie African Diaspora Fellowship Program (CADFP).

## References

1. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv:2108.07258 (2021). <https://doi.org/10.48550/arXiv.2108.07258>
2. Codella, F.N.C., et al.: MedImageInsight: An Open-Source Embedding Model for General Domain Medical Imaging. arXiv:2410.06542 (2024). <https://doi.org/10.48550/arXiv.2410.06542>
3. Gichoya, J.W.: AI for African Healthcare: Beyond Chatbots Toward Contextually Informed Innovation. Cassyni Seminar (2025), <https://doi.org/10.52843/cassyni.bsypvn>, Accessed 10 August 2025

4. Johnson, A., Pollard, T., Mark, R., Berkowitz, S., Horng, S.: MIMIC-CXR Database (version 2.1.0). *PhysioNet* (2024). <https://doi.org/10.13026/4jqj-jw95>, version 2.1.0, public chest X-ray dataset; RRID:SCR\_007345
5. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* **6**(1), Article 317 (2019). <https://doi.org/10.1038/s41597-019-0322-0>
6. Kotter, E., D'Antonoli, T.A., Cuocolo, R., Hierath, M., Huisman, M., Klontzas, M.E., Martí-Bonmatí, L., May, M.S., Neri, E., Nikolaou, K., et al.: Guiding AI in radiology: ESR's recommendations for effective implementation of the European AI Act. *Insights into Imaging* **16**(1), 33 (2025). <https://doi.org/10.1186/s13244-025-01905-x>
7. Lawrence, R., Dodsworth, E., Massou, E., Sherlaw-Johnson, C., Ramsay, A.I., Walton, H., O'Regan, T., Gleeson, F., Crellin, N., Herbert, K., et al.: Artificial intelligence for diagnostics in radiology practice: a rapid systematic scoping review. *eClinicalMedicine* **83** (2025). <https://doi.org/10.1016/j.eclinm.2025.103228>
8. Li, W., Li, L., Xiang, T., Liu, X., Deng, W., Garcia, N.: Can multiple-choice questions really be useful in detecting the abilities of LLMs? *arXiv:2403.17752* (2024). <https://doi.org/10.48550/arXiv.2403.17752>
9. Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J.: Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering* **35**(1), 857–876 (2021). <https://doi.org/10.1109/TKDE.2021.3090866>
10. Marivate, V.: DSFSI+AfriDSAI Hosts Prof Judy Wawira Gichoya. *Substack Newsletter* (2025), <https://dsup.substack.com/p/dsfsiafridsai-hosts-prof-judy-wawira>, accessed 10 August 2025
11. Moyo, S., Ismail, F., Van der Walt, M., Ismail, N., Mkhondo, N., Dlamini, S., Mthiyane, T., Chikovore, J., Oladimeji, O., Mametja, D., et al.: Prevalence of bacteriologically confirmed pulmonary tuberculosis in South Africa, 2017–19: a multistage, cluster-based, cross-sectional survey. *The Lancet Infectious Diseases* **22**(8), 1172–1180 (2022). [https://doi.org/10.1016/S1473-3099\(22\)00149-9](https://doi.org/10.1016/S1473-3099(22)00149-9)
12. Munyangaju, I., José, B., Bassat, Q., Esmail, R., Tlhapi, L.H., Maphophe, M., Mutemba, C., Cossa, L., Perez, P., Palmer, M., et al.: Assessment of radiological capacity and disparities in tb diagnosis: a comparative study of mozambique, south africa and spain. *BMJ Public Health* **2**(2) (2024). <https://doi.org/10.1136/bmjph-2024-001392>
13. Olatunji, T., Nimo, C., Owodunni, A., Abdullahi, T., Ayodele, E., Sanni, M., Aka, C., Omofoye, F., Yuehgoh, F., Faniran, T., et al.: AfriMed-QA: a Pan-African, multi-specialty, medical question-answering benchmark dataset. *arXiv:2411.15640* (2024). <https://doi.org/10.48550/arXiv.2411.15640>
14. Pal, A., Lee, J.O., Zhang, X., Sankarasubbu, M., Roh, S., Kim, W.J., Lee, M., Rajpurkar, P.: ReXVQA: A Large-scale Visual Question Answering Benchmark for Generalist Chest X-ray Understanding. *arXiv:2506.04353* (2025). <https://doi.org/10.48550/arXiv.2506.04353>
15. Saeidnia, H.R., Nilashi, M.: From MYCIN to MedGemma: A Historical and Comparative Analysis of Healthcare AI Evolution. *InfoScience Trends* **2**(6), 18–28 (2025). <https://doi.org/10.61186/ist.202502.06.02>
16. Siam, M.K., Varela, A., IED, C.M.C., Faruk, M.J.H., Cheng, J.Q., Gu, H., Al Maruf, A., Aung, Z.: Benchmarking Large Language Models on USMLE: Evaluating ChatGPT, DeepSeek, Grok, and Qwen in Clinical Reasoning and Medical Licensing Scenarios. *Research Square* (2025). <https://doi.org/10.21203/rs.3.rs-6651111/v1>

## Appendix A. Example Workflow

The following is an example of a report that could be generated by applying the RAISE methodology. The example image can be seen in figure 4 and it was taken from the MIMIC-CXR database [5]. Each section of the report starts with a heading of the following format: Level - SAQ - Answer (Confidence). Only one SAQ is shown per level in order to demonstrate the example and the redacted questions are still shown for each level.

### Structured Assessment Report

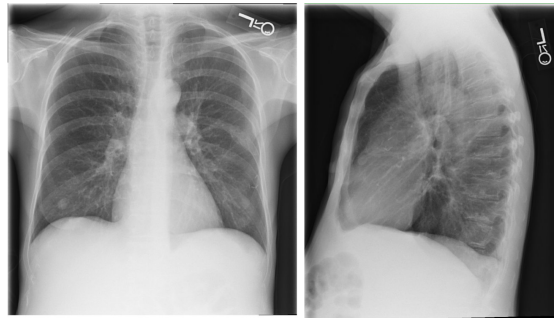


Fig. 4. Chest X-Ray image from the MIMIC-CXR database [4]

#### Level 0 – Context – Chest X-ray (95%)

These are clearly medical radiographic images of the chest cavity, showing the characteristic black and white contrast of X-ray imaging. The left image shows a frontal (posteroanterior) view displaying both lungs, ribcage, heart silhouette, and spine, while the right image shows a lateral (side) view of the same anatomical structures. The formatting, image quality, and anatomical structures visible are consistent with standard chest X-ray imaging used for diagnostic purposes in medical settings.

#### Level 1 – Patient Positioned Correctly – Yes (85%)

The PA view shows good symmetry with the clavicles appearing equidistant from the spine, indicating proper centering and rotation. The scapulae appear adequately moved out of the lung fields. The lateral view shows appropriate overlap of the ribs posteriorly and clear visualization of the retrocardiac space. Both views demonstrate adequate inspiration with good lung expansion, suggesting the patient was positioned correctly for standard chest radiography.

Other questions for level 1 are: “Exposure is Adequate?” and “No or Acceptable Blurring?”.

#### Level 2 – Is there a pleural effusion present? – No (85%)

Both costophrenic angles appear sharp and clear without the characteristic

blunting or meniscus sign that would indicate pleural fluid accumulation. The lateral view also shows clear posterior costophrenic angles. There is no evidence of fluid layering or obscuration of the hemidiaphragms that would suggest pleural effusion. The lung bases appear well-aerated and extend normally to the chest wall, indicating absence of significant pleural fluid collection.

Other question for level 2: “Are the opacities diffuse?”

**Level 3 – Is there an abnormal opacity in the upper lobes? – Yes (92%)**

There is a clearly visible abnormal opacity in the right upper lobe that appears as a dense, well-defined mass or consolidation. This opacity is distinctly different from the normal lung parenchyma and vascular markings, appearing more radiopaque than the surrounding lung tissue. The opacity is prominent on both the PA and lateral views, confirming its presence in the upper lobe region and indicating a pathological process that requires further clinical evaluation.

Other questions for level 3 are: “Is there evidence of cavitation in the affected areas?”, “Is there hilar or mediastinal lymphadenopathy?” and “Is the pathology severe?”.

**Level 4 – Do the radiographic findings suggest pulmonary tuberculosis as the most probable diagnosis? – Yes (78%)**

The radiographic findings of an upper lobe opacity/consolidation without obvious cavitation are consistent with pulmonary tuberculosis, which classically affects the upper lobes and apical segments of lower lobes. The location and appearance of the abnormality, combined with the absence of clear hilar lymphadenopathy (which is more common in primary TB in children), suggests post-primary or reactivation tuberculosis. However, other conditions like lung cancer, pneumonia, or other granulomatous diseases could present similarly, so clinical correlation, sputum studies, and possibly further imaging would be needed for definitive diagnosis.

Other question for level 4: “Given the image alone, is TB more likely than alternative causes (e.g., bacterial pneumonia, lung cancer)?”

**Original Label for data:**

A region of heterogeneous consolidation in the left upper lobe extends from the mediastinum posteriorly nearly to the chest wall. There may also be retrotracheal adenopathy. Although this could be any bacterial pneumonia, it raises serious concern for acute tuberculosis, particularly if the patient is on immunosuppressant medication. Lower lungs are clear. There is no pleural effusion. Heart size normal.

**Radiologist Assessment:**

The image represents a normal CXR without findings of Tuberculosis. Post surgical changes in the left chest. Based on RAISE, the radiologist rejected Level 3 and 4, and accepting performance at Level 2.