The Intention-Execution Disconnect in Medical AI: The ReXecution Framework for Evaluating Real-World Clinical Performance

Oishi Banerjee[1*], Lucas Bijnens[1,2*], Subathra Adithan[3] and Pranav Rajpurkar[1†]

[1] *Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA*
[2] *Department of Imaging and Pathology, KU Leuven, Leuven, Belgium*
[3] *Department of Radiodiagnosis, Jawaharlal Institute of Postgraduate Medical Education and Research, Puducherry, India*
[†] *Corresponding Author E-mail: pranav_rajpurkar@hms.harvard.edu*

[*] Both authors contributed equally to this work.

Abstract: We present the ReXecution framework for conducting clinician-centered assessments of medical AI assistants, providing detailed insights into their reliability in realistic clinical settings. Using this framework, we assessed AI assistants for chest X-ray (CXR) interpretation, exploring the gap between current model capabilities and real-world radiological needs. Unlike prior benchmarks that rely on automatically generated questions with limited clinical relevance, our dataset consists of 100 expert-curated tasks that radiologists might realistically present to an AI assistant in their day-to-day workflow. Through detailed manual review by a radiologist, we evaluated two leading foundation models, ChatGPT-o3 and MedGemma, on our tasks. While both models demonstrated considerable medical knowledge and reasoning capabilities on our tasks, they frequently struggled to interpret images and execute tasks accurately, producing correct outputs in only 5-10% of cases. Our detailed manual evaluation highlights a critical mismatch: models often abstractly understand radiology concepts but cannot reliably execute their plans when interpreting specific medical images. This work identifies key gaps in current models' ability to serve as comprehensive radiology assistants and provides insights into how the development and evaluation of models can better align with real-world clinician needs, enabling seamless clinician-AI collaboration.

Keywords: Clinician-AI Collaboration, Radiologist-Centered Evaluation, Vision-Language Models, AI for Chest X-Rays

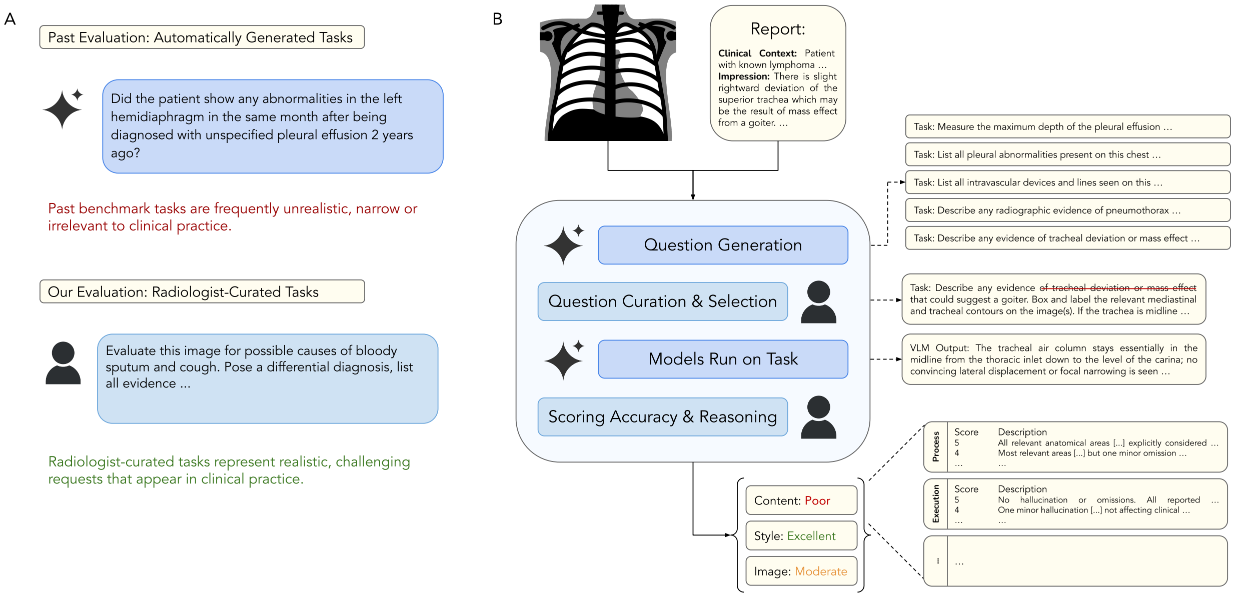
1. Introduction

Artificial intelligence (AI) for medicine has rapidly progressed from narrow classifiers to large-scale vision-language models (VLMs) that promise to function as generalist assistants for clinicians.[1,2] In radiology, some recent studies have found that AI foundation models can match or even exceed clinician performance on chest X-ray (CXR) interpretation tasks.[3] Such VLMs may collaborate with clinicians as interactive chatbots, providing measurements and other useful information as radiologists interpret a study.[4] Alternatively, models might perform useful tasks in the background before a clinician even reaches a study, flagging urgent cases or preemptively commenting on images according to a user's standing instructions.[5]

Several large-scale, automatically generated benchmarks have been constructed to evaluate assistive models on CXR tasks,[3,6,7] but these benchmarks prioritize scalability and automated scoring over clinical realism. Most are constructed from templates or report-mined labels, producing short-answer or multiple-choice tasks with clear-cut ground truths. While useful for coverage and automated scoring, such benchmarks overlook additional information encoded in the images, feature frequently unrealistic questions and answers, and ultimately diverge from how radiologists would actually interact with an AI assistant.

For instance, some benchmark questions are too trivial to require large VLMs (e.g., inferring patient sex from the image), while others are overly specific or complicated (e.g., "Q: Does the left mid lung zone contain either low lung volumes or calcified nodule? A: Yes.").[6] Moreover, such short ground-truth answers can obscure whether a model arrived at the correct response through accurate reasoning, chance, or shortcuts. The resulting benchmarks do not accurately represent the open-ended questions and complex reasoning processes that arise in clinical practice (Figure 1a).

To address these shortcomings, clinician-centered evaluation procedures are needed. In this exploratory study, we evaluate the capabilities of cutting-edge vision-language models, testing them on a complex, clinically relevant dataset created through radiologist-AI collaboration. We score model outputs on several dimensions, assessing both their "intentions"—their general medical knowledge and abstract planning capabilities—as well as their "execution" when interpreting specific images and making practical judgments. Our data curation and scoring processes comprise the ReXecution (*Re*alistic Intention-*Execution* Assessment) framework (Figure 1b), which can be extended beyond radiology to conduct clinician-centered assessments in other domains of medicine.


A
Past Evaluation: Automatically Generated Tasks
Our Evaluation: Radiologist-Curated Tasks
B
Report:
Question Generation
Question Curation & Selection
Models Run on Task
Scoring Accuracy & Reasoning
Content: Poor
Style: Excellent
Image: Moderate
Process
Execution
Score
Description
⋮

Did the patient show any abnormalities in the left hemidiaphragm in the same month after being diagnosed with unspecified pleural effusion 2 years ago?

Past benchmark tasks are frequently unrealistic, narrow or irrelevant to clinical practice.

Evaluate this image for possible causes of bloody sputum and cough. Pose a differential diagnosis, list all evidence ...

Radiologist-curated tasks represent realistic, challenging requests that appear in clinical practice.

Clinical Context: Patient with known lymphoma ... **Impression:** There is slight rightward deviation of the superior trachea which may be the result of mass effect from a goiter. ...

Task: Measure the maximum depth of the pleural effusion ...

Task: List all pleural abnormalities present on this chest ...

Task: List all intravascular devices and lines seen on this ...

Task: Describe any radiographic evidence of pneumothorax ...

Task: Describe any evidence of tracheal deviation or mass effect ...

Task: Describe any evidence ~~of tracheal deviation or mass effect~~ that could suggest a goiter. Box and label the relevant mediastinal and tracheal contours on the image(s). If the trachea is midline ...

VLM Output: The tracheal air column stays essentially in the midline from the thoracic inlet down to the level of the carina; no convincing lateral displacement or focal narrowing is seen ...

Score	Description
5	All relevant anatomical areas [...] explicitly considered ...
4	Most relevant areas [...] but one minor omission ...
...	...

Score	Description
5	No hallucination or omissions. All reported ...
4	One minor hallucination [...] not affecting clinical ...
...	...

...

Fig. 1: The ReXecution framework emphasizes (a) realistic task design and (b) clinician-led data curation and scoring, providing detailed insights into how models reason in clinical settings. [a]

By applying this framework to chest X-ray interpretation, we make two primary contributions:

(1) We curated a high-quality dataset with 100 challenging, realistic evaluation tasks. Selected from a list of automatically-generated candidates, each task was reviewed and edited by a radiologist to ensure that it would plausibly occur in clinical practice—either when actively interacting with a chatbot or when specifying tasks for an agent to perform independently in the background.

(2) Using our dataset, we tested ChatGPT-o3, a leading general-purpose foundation model, and MedGemma, a top medical foundation model trained heavily on chest X-ray inter-

[a] Figure 1a's "past evaluation" question comes from an existing benchmark linking EHR and CXR data.[8] We removed a patient identification number but otherwise preserved the text.

pretation. As our tasks elicited lengthy, nuanced answers, we manually evaluated model performance. We obtained rich qualitative insights into radiologist experiences with AI assistants, while also compiling descriptive statistics that highlight model weaknesses.

(3) We found that both models demonstrate sophisticated medical knowledge, accurately understand complex questions, and plan multi-step strategies to answer them. However, they struggled to execute those strategies due to the "visual" component of visual question-answering, frequently misreading chest X-rays and ultimately arriving at incorrect conclusions. Our findings indicate that, despite achieving expert-level performance on certain tasks, AI models still fall short of providing truly comprehensive assistance to clinicians.

2. Related Works

Prior visual question-answering datasets for chest X-ray interpretation have typically mined large numbers of questions from radiology reports, weakly labeled bounding boxes, or other metadata. The recently released ReXVQA benchmark features approximately 696,000 LLM-generated multiple-choice questions based on radiology reports. On a subset of 200 questions, this benchmark found that MedGemma outperformed radiology residents with an accuracy of 83.8%, though its performance varied depending on the type of task.[3]

Building on an earlier dataset,[8] MIMIC-Ext-MIMIC-CXR-VQA generated approximately 377,000 CXR questions by filling templates using AI-generated bounding boxes and keywords extracted from radiology reports. About two-thirds of MIMIC-Ext-MIMIC-CXR-VQA questions require yes/no or multiple-choice answers, while the rest expect a short list of abnormalities or locations.[6] More recently, Medical-CXR-VQA and MIMIC-Diff-VQA each presented approximately 700,000 questions, generated through a similar methodology.[7,9]

The GEMeX benchmark contains about 1,600,000 LLM-generated questions based on radiology reports and bounding boxes, as well as a GPT-based scoring metric.[10] While GEMeX also focuses on multiple-choice questions or brief answers, it expands on prior benchmarks by accepting multimodal outputs with both text and bounding boxes and by including more of an emphasis on causal reasoning (e.g. "What implications does the tortuous aorta have?").

While these recent datasets have been automatically constructed, the earlier SLAKE and VQA-RAD datasets contain tens of thousands of clinician-written questions for several radiology modalities including chest X-rays. These questions also expect short, often one-word answers that can be easily be automatically scored.[11,12] Outside visual question-answering, there have been attempts to automatically score longer free-text responses in the field of radiology report generation, yet automated metrics have struggled to evaluate complex descriptions of chest X-rays, offering only moderate correlations with expert opinions.[13,14]

3. Methods

3.1. *ReXecution Data Curation*

The first step of our framework is the construction of a dataset that reflects real-world clinical workflows while leveraging LLMs for efficiency. To this end, we developed a pipeline (Figure 1) to generate high-level, clinically relevant tasks for CXR interpretation through clinician-AI collaboration. We randomly sampled 100 chest X-ray studies from the test set of *MIMIC-CXR-JPG*, a large public dataset of CXRs and associated reports from Beth Israel Deaconess Medical Center in Boston, MA.[15]

For each study, we prompted GPT-o4-mini (Supplementary Table 1) to propose five candidate tasks across three categories. We selected these categories to require complex clinical reasoning and the integration of multiple imaging findings:

(1) *Broad abnormality query.* “Describe all pleural abnormalities”, “List all visible tubes and lines and assess for complications”, etc.
(2) *High-level diagnostic reasoning.* “Gather evidence of congestive heart failure”, “Perform differential diagnosis for a lobar opacity”, etc.
(3) *Measurements and temporal comparisons.* “Measure the cardiothoracic ratio”, “Assess effusion size compared to prior”, etc.

We prompted the model to generate questions that were relevant to the indication or findings from each case’s radiology report. Indications are provided before performing the imaging study and describe the symptoms or other clinical motivation for the study. Findings are reported after the study is performed and represent a radiologist’s interpretation of the images. By drawing questions from not only the radiology report but also the clinical context, we cover multiple phases in the clinical workflow—both before and after the radiologist starts working on the imaging study.

After generating questions, we prompted GPT-o4-mini (Supplementary Table 1) to score all questions on 5-point Likert scales, based on whether it (i) is clinically relevant, (ii) is answerable from CXR alone, and (iii) requires reasoning beyond simple pattern matching. Additionally, the LLM assigned multiple types of content tags (e.g., areas of interest, main pathology). The tagging and scoring system are described in more detail in the appendix (Supplementary Table 2).

A radiologist then selected a final question for each case, considering high-scoring candidates first and re-scoring selected questions to verify that they aligned with our goals. During this process, they verified that selected questions were clinically important and clearly phrased. They also leveraged the content tags to ensure that the selected question pool covered a large variety of medical topics and tasks, reflecting realistic caseloads. Where needed, the radiologist rephrased questions or wrote new ones to maintain quality, diversity and clinical relevance. Additionally, while most questions (n=89) concerned a single case, the radiologist verified whether MIMIC-CXR contained a sufficiently recent prior study for questions covering disease progression over time.

The resulting tasks covered a wide range of clinical use cases. Based on the content tags, many areas of interest were represented: lung parenchyma (37%), pleura (15%), cardiac silhouette and mediastinum (14%), hilar regions (11%), devices (9%), osseous structures (4%), diaphragm and soft tissues (6%), and airways (3%). Pathology tags covered pneumonia, effusions, atelectasis, lymphadenopathy, pneumothorax as well as some rarer pathologies—aligning with the clinical distribution seen in routine radiology practice.

A score histogram (Supplementary Figure 1a) shows the distribution of quality scores across the selected question set, indicating the overall high quality of the included items. For more details on the variation in our dataset, breakdown plots are provided in the appendix (Supplementary Figure 1b).

3.2. *Vision-Language Models*

We evaluated two state-of-the-art VLMs that represent different development paradigms:

- **ChatGPT-o3 (OpenAI, closed-source generalist)**. The strongest reasoning model in the GPT family at time of writing. It is a multimodal foundation model capable of ingesting multiple images and producing both text and figures.
- **MedGemma-4b-it (Google, open-source medical)**. A vision-language model fine-tuned on medical imaging with strong performance on multiple-choice CXR benchmarks. We selected MedGemma specifically for its strong performance on ReXVQA, a multiple-choice benchmark where it was found to exceed the performance of radiology residents.[3] It is optimized for processing one input image at a time and returns text-only outputs.

For ChatGPT-o3, we provided all available CXR views and instructed the model to generate bounding-box annotations for localized findings. If no abnormalities were found, the model was not required to draw boxes; when ChatGPT-o3 found abnormalities but did not annotate them, we did not force it to generate bounding boxes. For MedGemma, we provided a single PA (or AP if unavailable) image per case, or both images only for temporal comparisons. No bounding boxes were requested as this is not supported by MedGemma. Except for the bounding box directions, the prompts were harmonized across the models to reduce bias.

3.3. *ReXecution Metrics*

In our framework, we assess text outputs on four dimensions (Figure 2) that capture both a model's abstract medical knowledge and concrete visual abilities. Each text output was scored by a radiologist on each dimension:

(1) *Process.* Did the model correctly interpret the task and plan an appropriate strategy? (e.g., while gathering evidence of cancer, but only looking for lung lesions without considering bone lesions)
(2) *Execution.* Did it accurately identify and describe image findings without hallucinations or omissions? (e.g., attempting to look for lung lesions, but hallucinating a granulomatous lesion)
(3) *Synthesis.* Did it integrate findings into sound, clinically appropriate conclusions, accounting for uncertainty? (e.g., claiming that a granuloma is suggestive of cancer while ignoring more plausible explanations such as infection)
(4) *Language.* Was the response clear, concise, and professional? (e.g., rambling and repeating itself when answering a question)

In practice, radiologists would struggle to use a model with poor Execution or Synthesis scores, and they would struggle to trust a model with poor Process scores even if it sometimes guessed at correct answers. We therefore calculate an aggregated *Content* score, which judges

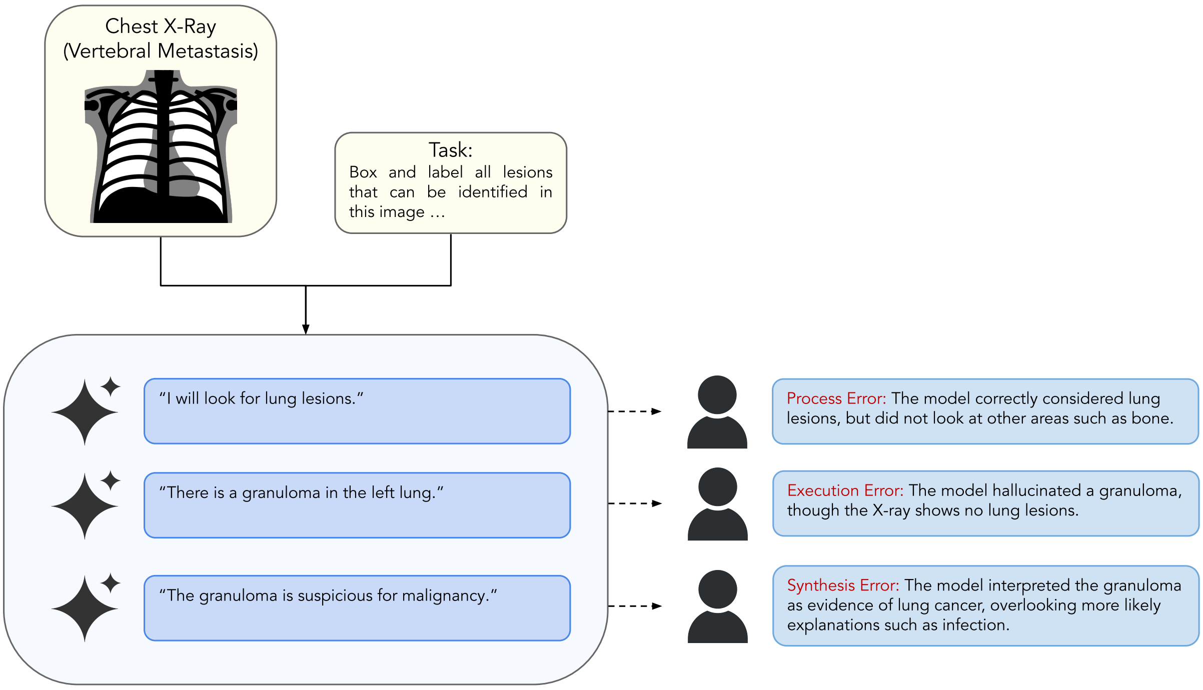

Chest X-Ray
(Vertebral Metastasis)
Task:
Box and label all lesions that can be identified in this image ...
"I will look for lung lesions."
"There is a granuloma in the left lung."
"The granuloma is suspicious for malignancy."
Process Error: The model correctly considered lung lesions, but did not look at other areas such as bone.
Execution Error: The model hallucinated a granuloma, though the X-ray shows no lung lesions.
Synthesis Error: The model interpreted the granuloma as evidence of lung cancer, overlooking more likely explanations such as infection.

Fig. 2: Demonstration of different content error types. The ReXecution framework judges content on Process, Execution, and Synthesis.

output content based on its weakest link. The Content score is equal to the minimum of the Process, Execution, and Synthesis scores. In other words, an output cannot receive a Content score of 5 unless it successfully performs all relevant tasks required to answer the question correctly—ranging from question comprehension, detection of findings, to reasoning about their clinical implications.

Additionally, image outputs from ChatGPT-o3 were scored on two dimensions by a radiologist. *Image Content* scores reflect whether the model correctly localizes the regions it is attempting to box. If a model attempts to label a pleural effusion but instead boxes a lower part of the hemidiaphragm, this error would be considered an Image Content issue. *Image Style* scores reflect whether the model plans out a useful illustration, attempting to box findings mentioned in the text without cluttering the image with unrelated boxes. If a model boxes and labels a pacemaker when asked to calculate the cardiothoracic ratio, this unnecessary addition would be considered an Image Style issue.

Scoring for each dimension followed a customized 5-point scale (1 = poor, 5 = excellent). The full rubric is provided in the appendix (Supplementary Table 2).

3.4. *Statistical Analysis*

For each of the five text scores, we tested whether there was a significant difference between ChatGPT-o3 and MedGemma by applying the Wilcoxon signed-rank test for paired non-parametric comparisons and controlling for multiple testing using the Benjamini–Hochberg false discovery rate procedure. We did not run statistical tests between different score types (e.g. Content vs. Language) because each dimension was measured using custom criteria, making formal statistical comparisons infeasible. Aggregate means are reported with standard deviations.

3.5. *Inter-Reader Agreement*

To gauge the reproducibility of our evaluation rubric, a senior radiologist independently rescored roughly a quarter of the dataset. They labeled 51 model outputs, 15 of which contained boxed images from ChatGPT-o3. Agreement with the primary reader was quantified using quadratic-weighted Cohen's κ (QWK), appropriate for our five-point ordinal scale, and mean absolute differences (MAD). Where both readers provided scores, we used the average of the two scores for our final results to provide more robust performance metrics.

4. Results

4.1. *Quantitative Overview of Text Content and Language*

We found that both models struggled to produce factually accurate answers (Table 1). ChatGPT-o3 received an average Content score of 2.71, with 10.0% of outputs receiving a 5. MedGemma received an average Content score of 2.84, with only 4.9% of outputs receiving a score of 5. MedGemma's failure to generate an answer for 19 tasks was concerning. The underlying issue was a built-in limit on the number of tokens it can process at once, which was triggered by tasks that elicited particularly detailed answers. However, both models generally maintained a professional writing style and used appropriate radiological vocabulary, with 56.8% of MedGemma outputs and 31.0% of ChatGPT-o3 outputs receiving Language scores of 5. When comparing ChatGPT-o3 and MedGemma's performance, we tested only data points that received outputs from both models and found no statistically significant difference.

Metric	ChatGPT-o3 (n = 100)	MedGemma (n = 81)	p_{adj}
Content	2.71 ± 1.31	2.84 ± 1.14	0.52
Language	4.15 ± 0.73	4.33 ± 0.96	0.66

Table 1: Aggregate metrics for content and language clarity.

4.2. *Breakdown of Content Errors*

Errors were frequent across all four content categories (Table 2), driving the low overall content scores seen above. We noted that both models received their highest average scores in Process and their lowest average scores in Execution, with gaps of .65 for MedGemma and 1.27 for ChatGPT-o3. Differences in error distributions also differed strikingly across categories (Figure 3). Only 5.0% of MedGemma's Process scores and 8.0% of ChatGPT-o3's Process scores fell under 3, compared to 34.5% of MedGemma's Execution scores and 42% of ChatGPT-o3's

Execution scores. Synthesis scores were left-skewed like Process scores, indicating that scores were rarely very poor. After correcting for multiple-hypothesis testing, we again found no statistically significant difference between ChatGPT-o3 and MedGemma.

Metric	ChatGPT-o3 (n = 100)	MedGemma (n = 81)	p_{adj}
Process	4.17 ± 0.95	3.91 ± 0.94	0.10
Execution	2.90 ± 1.44	3.26 ± 1.27	0.19
Synthesis	3.85 ± 1.06	3.49 ± 1.21	0.10

Table 2: Content metric breakdown in individual task performances.

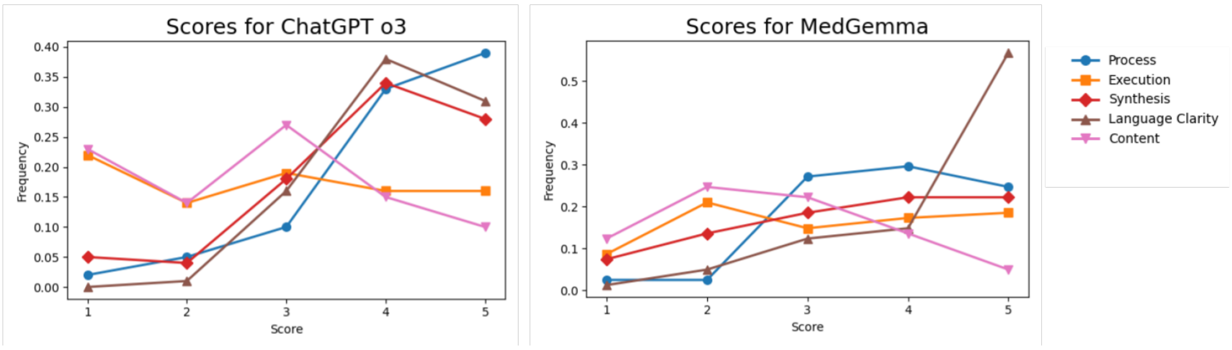


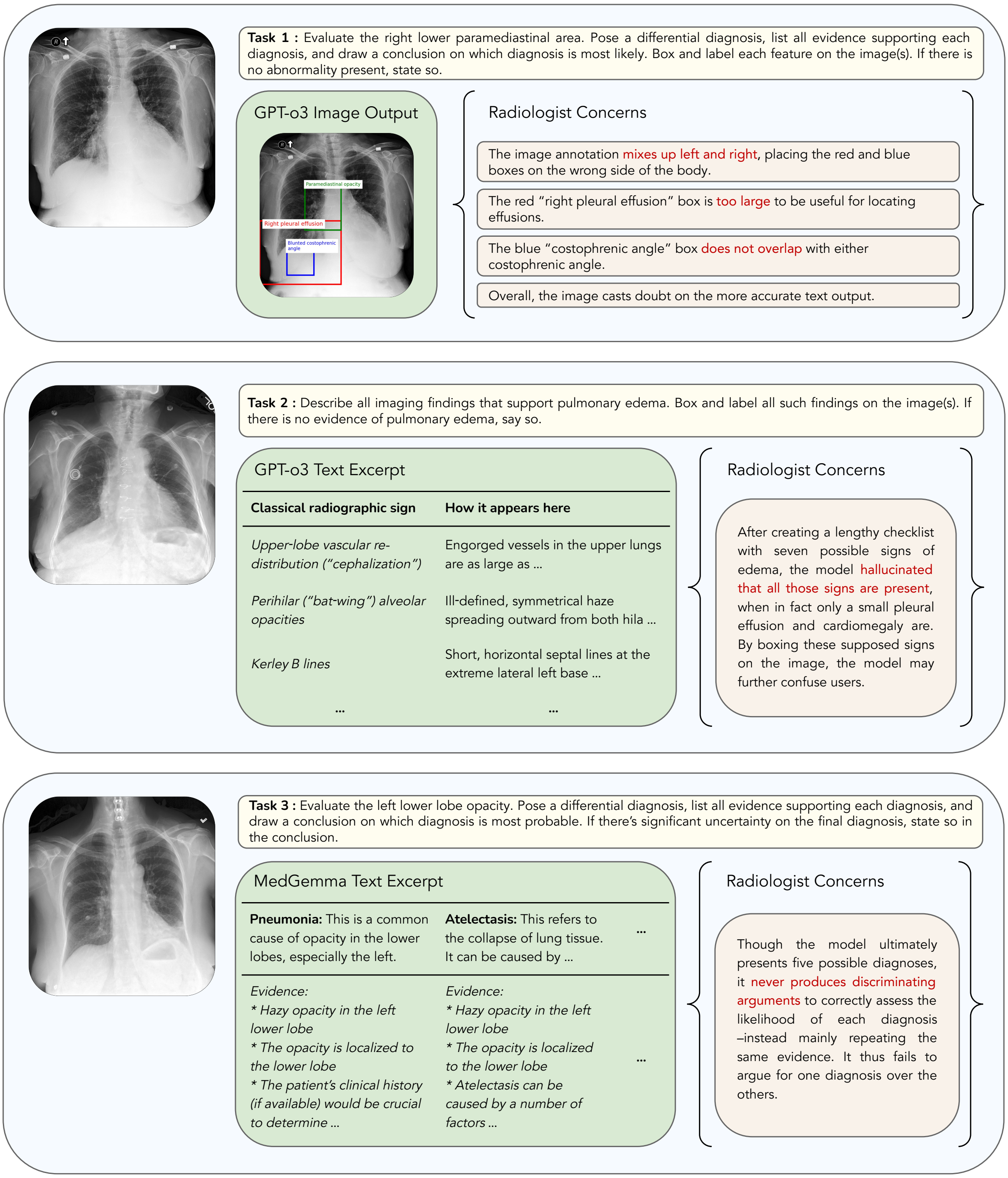
Fig. 3: Score distributions showing both models avoided low scores on Process but had a large spread in Execution scores and rarely reach an overall Content score of 5.

4.3. *Overview of ChatGPT-o3 Image Quality*

For 62 tasks, ChatGPT-o3 successfully produced images, otherwise deciding that no image was necessary or failing to follow that part of the instructions. Upon assessment of these images, we found frequent errors in ChatGPT-o3's boxes and labels, with an average Image Content score of 3.06±1.22. Only 10.1% of images received a 5 for Image Content, indicating that most images contained a misplaced box or label. ChatGPT-o3 performed better at Image Style with an average score of 3.63±1.33. 36.2% of images received an Image Style score of 5, indicating that ChatGPT-o3 generally showed better judgment when planning what to box and label.

4.4. *Qualitative Observations*

As suggested by the quantitative findings, we found that both models showed strong theoretical knowledge while struggling to interpret images correctly in practice, yet we also observed distinct behavioral patterns and failure modes across the two models (Figure 4). ChatGPT-o3 often found the correct regions of interest but misclassified the abnormality type. It was also prone to making mistakes on laterality, mixing up the left and right sides of the image; this



Task 1 : Evaluate the right lower paramediastinal area. Pose a differential diagnosis, list all evidence supporting each diagnosis, and draw a conclusion on which diagnosis is most likely. Box and label each feature on the image(s). If there is no abnormality present, state so.

GPT-o3 Image Output

Radiologist Concerns

The image annotation mixes up left and right, placing the red and blue boxes on the wrong side of the body.

The red "right pleural effusion" box is too large to be useful for locating effusions.

The blue "costophrenic angle" box does not overlap with either costophrenic angle.

Overall, the image casts doubt on the more accurate text output.

Task 2 : Describe all imaging findings that support pulmonary edema. Box and label all such findings on the image(s). If there is no evidence of pulmonary edema, say so.

GPT-o3 Text Excerpt

Classical radiographic sign	**How it appears here**
Upper-lobe vascular re-distribution ("cephalization")	Engorged vessels in the upper lungs are as large as ...
Perihilar ("bat-wing") alveolar opacities	Ill-defined, symmetrical haze spreading outward from both hila ...
Kerley B lines	Short, horizontal septal lines at the extreme lateral left base ...
...	...

Radiologist Concerns

After creating a lengthy checklist with seven possible signs of edema, the model hallucinated that all those signs are present, when in fact only a small pleural effusion and cardiomegaly are. By boxing these supposed signs on the image, the model may further confuse users.

Task 3 : Evaluate the left lower lobe opacity. Pose a differential diagnosis, list all evidence supporting each diagnosis, and draw a conclusion on which diagnosis is most probable. If there's significant uncertainty on the final diagnosis, state so in the conclusion.

MedGemma Text Excerpt

Pneumonia: This is a common cause of opacity in the lower lobes, especially the left.	**Atelectasis:** This refers to the collapse of lung tissue. It can be caused by ...	...
Evidence: ** Hazy opacity in the left lower lobe* ** The opacity is localized to the lower lobe* ** The patient's clinical history (if available) would be crucial to determine ...*	*Evidence:* ** Hazy opacity in the left lower lobe* ** The opacity is localized to the lower lobe* ** Atelectasis can be caused by a number of factors ...*	...

Radiologist Concerns

Though the model ultimately presents five possible diagnoses, it never produces discriminating arguments to correctly assess the likelihood of each diagnosis –instead mainly repeating the same evidence. It thus fails to argue for one diagnosis over the others.

Fig. 4: Error examples highlighting radiologist's concerns on three cases.

pattern has definite potential to mislead clinicians, who can also struggle with this aspect of radiological imaging. Often, ChatGPT-o3 tended to explain its reasoning in depth, even when arriving at incorrect conclusions. This verbosity aided transparency but occasionally reduced clarity, confusing clinicians or slowing them down.

On the other hand, we found that MedGemma tended toward brevity, providing little insight into its reasoning, providing insufficient detail and under-reporting secondary findings. For example, when asked to gather all evidence of a pathology, the model described some findings like pleural effusions and consolidation while neglecting other findings such as rib crowding or volume loss. It also gave overly broad descriptions of abnormalities, such as stating that an "opacity" is present without further specifying that the opacity was an "air-bronchogram." For 19 tasks, MedGemma failed to finish generating answers due to an excessive token count; it is possible that those answers would have contained the desired level of detail if they had been successfully generated.

4.5. *Inter-Reader Agreement*

Moderate agreement was obtained for Execution ($\kappa = 0.57$), and Image Style ($\kappa = 0.48$). We observed lower rates of agreement on other dimensions (Table 3). Low rates of agreement were likely driven in part by the lack of variance in our score distributions; for example, nearly all Process scores fell into the range of 3-5, leading to lower Cohen's κ values. We note that mean absolute differences were under 1 for all text content categories; even when their orderings across cases differed, reader scores for any individual case were typically close together, as seen on Figure 5.

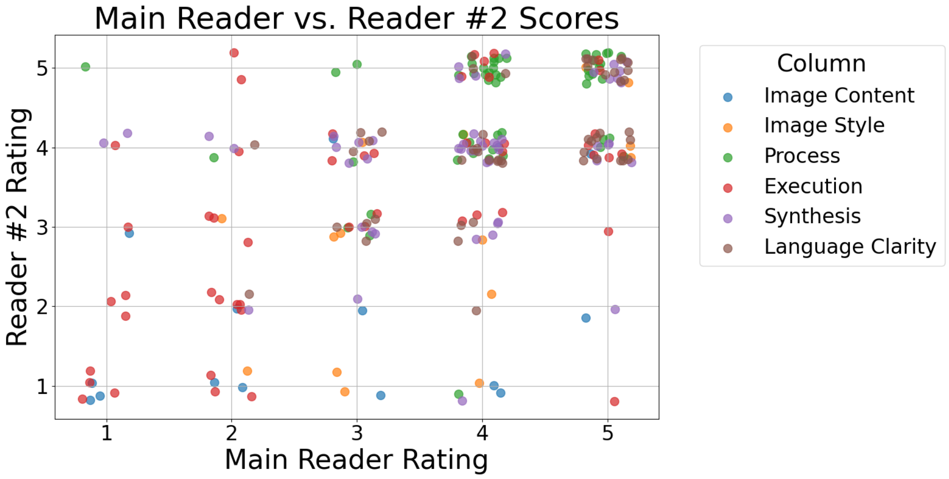


Fig. 5: Plot of both readers' scores, which generally differ by a maximum of one category.

5. Discussion

This study presented the ReXecution framework and evaluated the performance of two leading foundation models on expert-curated questions that reflect the interpretive demands that AI assistants would face in real-world radiology practice. Our evaluation set was built to mirror everyday radiological workstation questions requiring complex, detailed answers—contrasting with current benchmarks that rely on automatically generated multiple-choice or short-answer

Evaluation Category	QWK	MAD	Main Reader ($\mu \pm \sigma$)	Reader #2 ($\mu \pm \sigma$)
Process	0.215	0.667	4.118 ± 0.840	4.510 ± 0.784
Execution	0.573	0.882	2.961 ± 1.399	3.176 ± 1.337
Synthesis	0.291	0.706	3.824 ± 1.014	3.902 ± 0.878
Language	0.560	0.510	4.235 ± 0.862	4.078 ± 0.771
Image Content	0.327	1.200	2.733 ± 1.438	1.933 ± 1.223
Image Style	0.484	1.000	3.667 ± 1.047	2.933 ± 1.438

Table 3: Quadratic Weighted Kappa (QWK), Mean Absolute Difference (MAD), and reader score distributions (mean ± SD) for each evaluation category.

questions and fall short of modeling clinical settings. Through expert review of tasks and model answers, we gathered rich qualitative information on how current AI models would perform when supporting radiologists at the workstation.

Our findings reveal important insights into both the promise and the limitations of present-day foundation models in radiology. When performing our tasks, both ChatGPT-o3 and MedGemma displayed strong medical knowledge and abstract reasoning, with ChatGPT-o3's detailed explanations proving particularly useful. However, both struggled to execute their plans due to weaker image understanding—they often knew what to do, but not how to do it. Quantitatively, Process scores topped Execution scores by 0.65 points for MedGemma and 1.27 points for ChatGPT-o3, underscoring a gap between planning and pixel-level perception. Average Content scores were only 2.7-2.8 out of 5, and $\leq$ 10% of answers were marked as completely correct. In addition to observing errors in text outputs, we noted frequent errors in ChatGPT-o3's image outputs. Only 10.1% of 62 annotated images showed perfect localization, while nearly half of the figures were stylistically sound but spatially inaccurate. Ideally, image outputs would improve explainability and help users find subtle findings mentioned by text outputs, yet current image annotations are as likely to decrease trust in correct text outputs or confuse users.

We found no statistically significant differences between ChatGPT-o3, a leading general-purpose model, and MedGemma, a cutting-edge foundation model trained specifically for medical imaging tasks. While the two models demonstrated qualitatively different patterns of behavior, both impressed readers with their abstract medical knowledge regardless of their mistakes and occasional unreliable conclusions.

We observed interesting practical differences when using models, beyond the fact that only ChatGPT-o3 could illustrate its outputs. MedGemma was constrained in its inputs, officially recommending that users only provide one image at a time. This constraint substantially restricts the amount of relevant information users can provide, since imaging studies routinely contain chest X-rays taken from different views. MedGemma was also limited in the number of tokens it could process at once and thus failed to perform tasks that elicited longer, more complex outputs. To keep up with the offerings of large, general-purpose models, medical foundation models must make use of multimodal data and increasing amounts of context. The main usability issue we encountered with ChatGPT-o3 was that it sometimes failed to follow the "bounding box" parts of our instructions, outputting only a text output. In some instances, it completely ignored that part of our instruction, while in others it acknowledged the request but requested guidance on what to draw. While this behavior might be acceptable from an interactive chatbot, it could also prove frustrating in situations where models are expected to act independently and follow instructions the first time.

Broadly speaking, our findings complicate recent claims that chest X-ray models are reaching expert-level performance on classification and even radiology report generation. For example, a recent multiple-choice benchmark found that MedGemma achieved 83.8% accuracy on multiple-choice chest X-ray questions,[3] yet only 4.9% of MedGemma's outputs were free of content errors on our complex tasks. With a focus on radiologist-centered evaluation, our work shows the importance of "stress-testing" models on difficult, clinically relevant tasks, revealing shortcomings in model performance before they enter real-world workflows. More work is needed to help models match their theoretical prowess with concrete skills, perhaps by focusing on fine-grained image localization tasks and on findings like rib crowding or air bronchograms that—though common—may be overlooked by existing labeled datasets. Such advances can bridge the gap between current AI capabilities and real-world clinical needs, allowing models to provide comprehensive support and achieve true clinician-AI collaboration.

6. Limitations

While we cover a large range of tasks related to abnormality detection, diagnostic support, and measurement, we omitted some clinically relevant use cases, such as "quality control" tasks for flagging otherwise low-quality chest X-rays. We also did not include "unanswerable" questions that required models to request more information or refrain from answering. Since our focus was on posing realistic questions, we did not design questions to purposefully mislead the model by emphasizing irrelevant or counterintuitive details, though such "trick" questions may have revealed important biases and merit consideration in other work.

We occasionally encountered ambiguous edge cases when scoring answer content, as it is not always clear whether an error arises from the "Process", "Execution", or "Synthesis" phase; we attempted to mitigate this issue by aggregating those scores in a single "Content" score to penalize all content-related errors on a task, no matter the phase. Score subjectivity was also reflected in the fair-to-moderate inter-reader agreement—demonstrating inherent subjectivity in clinician evaluation and emphasizing the need for multi-reader evaluation. Finally, our semi-automated pipeline required considerable radiologist effort, even with only 100 studies. Future work may leverage LLMs more heavily, improving scalability while maintaining realism.

7. Supplementary Materials

Materials are available at https://rajpurkarlab.github.io/rexecution-supplementals/.

8. Acknowledgments

We thank Sung Eun Kim and Xiaoman Zhang for feedback. This work was supported by Biswas Family Foundation's Transformative Computational Biology Grant in collaboration with the Milken Institute. L.B. was supported by Junior Orsi and Barco.

References

1. A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, S. Xu, F. Jamil, C. Hughes, C. Lau, J. Chen, F. Mahvar, L. Yatziv, T. Chen, B. Sterling, S. A. Baby, S. M. Baby, J. Lai, S. Schmidgall, L. Yang, K. Chen, P. Bjornsson, S. Reddy, R. Brush, K. Philbrick, M. Asiedu, I. Mezerreg, H. Hu, H. Yang, R. Tiwari, S. Jansen, P. Singh, Y. Liu, S. Azizi, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Riviere, L. Rouillard, T. Mesnard, G. Cideron, J. bastien Grill, S. Ramos, E. Yvinec, M. Casbon, E. Buchatskaya, J.-B. Alayrac, D. Lepikhin, V. Feinberg, S. Borgeaud, A. Andreev, C. Hardin, R. Dadashi, L. Hussenot, A. Joulin, O. Bachem, Y. Matias, K. Chou, A. Hassidim, K. Goel, C. Farabet, J. Barral, T. Warkentin, J. Shlens, D. Fleet, V. Cotruta, O. Sanseviero, G. Martins, P. Kirk, A. Rao, S. Shetty, D. F. Steiner, C. Kirmizibayrak, R. Pilgrim, D. Golden and L. Yang, Medgemma technical report (2025).
2. H.-Y. Zhou, J. N. Acosta, S. Adithan, S. Datta, E. J. Topol and P. Rajpurkar, Medversa: A generalist foundation model for medical image interpretation (2025).
3. A. Pal, J.-O. Lee, X. Zhang, M. Sankarasubbu, S. Roh, W. J. Kim, M. Lee and P. Rajpurkar, Rexvqa: A large-scale visual question answering benchmark for generalist chest x-ray understanding (2025).
4. M. Moor, O. Banerjee, Z. Abad, H. Krumholz, J. Leskovec, E. Topol and P. Rajpurkar, Foundation models for generalist medical artificial intelligence, *Nature* **616**, 259 (04 2023).
5. J. Huang, M. T. Wittbrodt, C. N. Teague, E. Karl, G. Galal, M. Thompson, A. Chapa, M.-L. Chiu, B. Herynk, R. Linchangco, A. Serhal, J. A. Heller, S. F. Abboud and M. Etemadi, Efficiency and quality of generative ai–assisted radiograph reporting, *JAMA Network Open* **8**, e2513921 (06 2025).
6. S. Bae, D. Kyung, J. Ryu, E. Cho, G. Lee, S. Kweon, J. Oh, L. Ji, E. Chang, T. Kim and E. Choi, Mimic-ext-mimic-cxr-vqa: A complex, diverse, and large-scale visual question answering dataset for chest x-ray images `https://physionet.org/content/mimic-cxr-vqa/1.0.0/`, (2024), Published July 19, 2024.
7. X. Hu, L. Gu, K. Kobayashi, L. Liu, M. Zhang, T. Harada, R. Summers and Y. Zhu, Medical-cxr-vqa dataset: A large-scale llm-enhanced medical dataset for visual question answering on chest x-ray images `https://example.com/medical-cxr-vqa`, (2025), Published January 21, 2025.
8. S. Bae, D. Kyung, J. Ryu, E. Cho, G. Lee, S. Kweon, J. Oh, L. Ji, E. I.-C. Chang, T. Kim and E. Choi, Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images (2023).
9. X. Hu, L. Gu, Q. An, M. Zhang, L. Liu, K. Kobayashi, T. Harada, R. Summers and Y. Zhu, Medical-diff-vqa: A large-scale medical dataset for difference visual question answering on chest x-ray images `https://example.com/medical-diff-vqa`, (2025), Published February 3, 2025.
10. B. Liu, K. Zou, L. Zhan, Z. Lu, X. Dong, Y. Chen, C. Xie, J. Cao, X.-M. Wu and H. Fu, Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis (2025).
11. B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang and X.-M. Wu, Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering (2021).
12. J. J. Lau, S. Gayen, A. Ben Abacha and D. Demner-Fushman, A dataset of clinically generated visual questions and answers about radiology images, *Scientific Data* **5** (2018), Published November 20, 2018.
13. O. Banerjee, A. Saenz, K. Wu, W. Clements, A. Zia, D. Buensalido, H. Kavnoudias, A. S. Abi-Ghanem, N. E. Ghawi, C. Luna, P. Castillo, K. Al-Surimi, R. A. Daghistani, Y.-M. Chen, H. sheng Chao, L. Heiliger, M. Kim, J. Haubold, F. Jonske and P. Rajpurkar, Rexamine-global: A framework for uncovering inconsistencies in radiology report generation metrics (2024).
14. F. Yu, M. Endo, R. Krishnan, C. P. Langlotz, V. K. Venugopal and P. Rajpurkar, Evaluating

progress in automatic chest x-ray radiology report generation, *Patterns* **4**, p. 100802 (2023), Published September 8, 2023.

15. A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C. ying Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz and S. Horng, Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs (2019).