ReXVQA: A Large-scale Visual Question Answering Benchmark for Generalist Chest X-ray Understanding

Ankit Pal[1,3], Jung-Oh Lee[2], Xiaoman Zhang[3], Malaikannan Sankarasubbu[1],
Seunghyeon Roh[2], Won Jung Kim[2], Meesun Lee[2], Pranav Rajpurkar[3]

[1] *Saama AI Research, Saama Technologies, India*
E-mail: {ankit.pal, malaikannan.sankarasubbu}@saama.com

[2] *Seoul National University, Seoul, South Korea*
E-mail: {pisceanoh, seunghyeon.roh, wonjung.kim, meesun.lee}@snu.ac.kr

[3] *Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA*
E-mail: {xiaoman.zhang, pranav.rajpurkar}@hms.harvard.edu

🤗 **Dataset**	hf.co/datasets/rajpurkarlab/ReXVQA
Supplementary Material	Github Appendix.pdf

We present ReXVQA, the largest and most comprehensive benchmark for visual question answering (VQA) in chest radiology, comprising 694,841 questions paired with 160,000 chest X-rays studies across training, validation, and test sets. Unlike prior efforts that rely heavily on template based queries, ReXVQA introduces a diverse and clinically authentic task suite reflecting five core radiological reasoning skills: presence assessment, location analysis, negation detection, differential diagnosis, and geometric reasoning. We evaluate eight state-of-the-art multimodal large language models, including MedGemma-4B-it, Qwen2.5-VL, Janus-Pro-7B, and Eagle2-9B. The best-performing model (MedGemma) achieves 83.24% overall accuracy. To bridge the gap between AI performance and clinical expertise, we conducted a comprehensive human reader study involving 3 senior radiology residents on 200 randomly sampled cases. Our evaluation demonstrates that MedGemma achieved superior performance (83.84% accuracy) compared to human readers (best radiology resident: 77.27%), representing a significant milestone where AI performance exceeds human evaluation on chest X-ray interpretation. The reader study reveals distinct performance patterns between AI models and radiology residents, with strong inter-reader agreement among the human readers while showing more variable agreement patterns between human readers and AI models. ReXVQA establishes a new standard for evaluating generalist radiological AI systems, offering public leaderboards, fine-grained evaluation splits, structured explanations, and category-level breakdowns. This benchmark lays the foundation for next-generation AI systems capable of mimicking expert-level clinical reasoning beyond narrow pathology classification.

1. Introduction

Chest X-ray (CXR) interpretation requires a radiologist to perform diverse cognitive tasks - from localizing findings *where is the reticular opacity?* to comparative analysis *has the hilar* *enlargement progressed?* to offering differential diagnoses *what are the likely causes of these peripheral findings?* A truly generalist CXR AI system would need similar capabilities: flexibly answering questions about location, relationships, measurements, and diagnostic reasoning rather than just detecting predefined pathologies.

Current CXR AI approaches, while impressive at disease classification, operate within narrow constraints. Systems have progressed from detecting a handful of conditions to impressive performance on multi-label classification of up to 130 pathologies, achieving near-radiologist performance on specific tasks.[14,16,18] However, they remain fundamentally limited to a fixed set of predetermined labels and cannot engage in the broader analytical reasoning that characterizes expert radiological assessment.

The emergence of multimodal Large Language Models (LLMs) offers a promising path toward such generalist medical AI systems.[12] These models can process both images and natural language, potentially enabling them to engage in the kind of flexible visual reasoning and natural dialogue that characterizes clinical practice. Early results show these models can understand basic medical concepts and engage in simple diagnostic reasoning when prompted with medical images.[13,15] However, systematically evaluating these models' capabilities across clinically meaningful tasks remains challenging. While recent datasets have scaled in size and scope, most rely on templated question generation and lack the diversity and complexity of real clinical reasoning, limiting their effectiveness as generalist benchmarks.

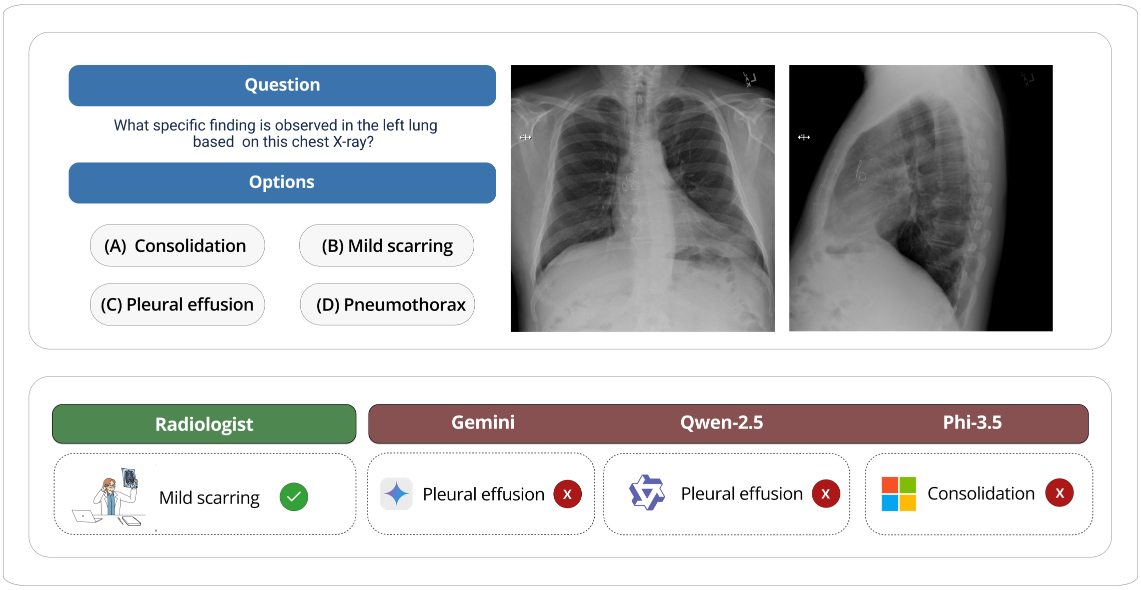

Fig. 1. Sample from the ReXVQA dataset, where human readers correctly identified mild scarring in the left lung base (correct answer B), while three state-of-the-art LVMs (Gemini, Qwen-2.5, and Phi-3.5) provided incorrect assessments, misidentifying the condition as pleural effusion or consolidation.

To address these limitations, we introduce ReXVQA, a benchmark of approximately 694,841 multiple-choice questions (MCQs) questions paired with 160,000 chest X-rays sourced from four U.S. health systems. Unlike previous datasets, ReXVQA evaluates five distinct cog-

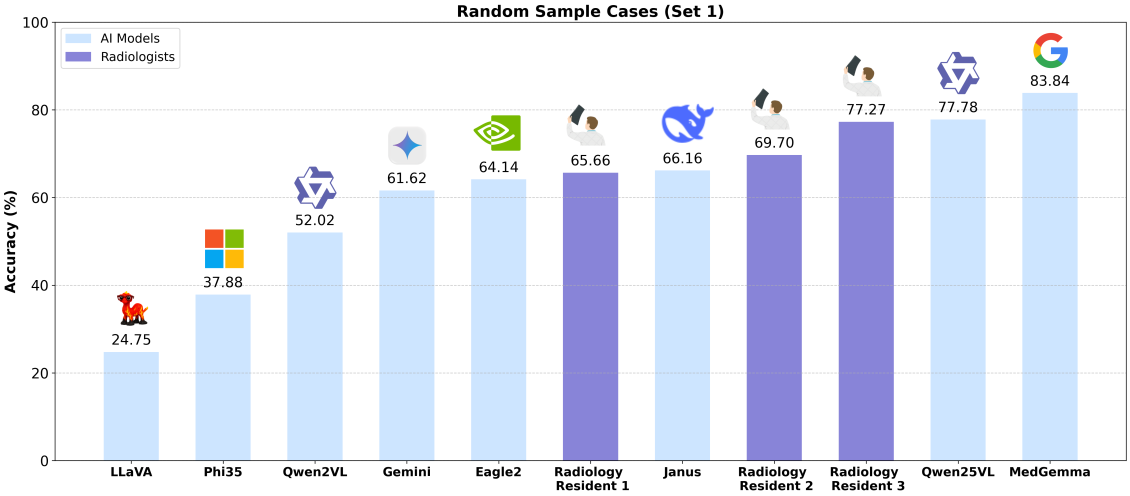

Fig. 2. Performance comparison of AI models and human readers across 200 random sampled cases. The bar chart shows overall accuracy (%) for eight AI models (Eagle2, Gemini, Janus, LLaVA, Phi35, Qwen2VL, Qwen25VL, and MedGemma) and three senior radiology residents.

nitive abilities that mirror clinical workflows, with questions distributed as follows: negation assessment (36.5% of questions), presence assessment (36.1%), differential diagnosis (20.9%), location and distribution assessment (6.1%), and geometric information analysis (0.4%). Questions are generated through a rigorous three-layer pipeline with expert-refined prompts developed over multiple rounds of radiologist feedback, ensuring they reflect authentic clinical reasoning patterns rather than artificial templates. Figure 1 shows one random sample from the dataset.

Our evaluation of eight state-of-the-art multimodal LLMs reveals significant advances in medical AI capabilities, with MedGemma demonstrating exceptional performance across all radiological reasoning tasks. MedGemma achieves superior performance in negation assessment (85.03%), presence assessment (85.21%), and location and distribution assessment (83.47%). The model shows remarkable capabilities across anatomical structures, achieving 91.84% on rib detection, 97.03% on heart findings, and 92.68% on spine assessment. Our reader study demonstrates a significant milestone: MedGemma surpasses senior radiology residents' performance on randomly sampled cases (83.84% vs. best reader: 77.27%), representing the first instance where AI consistently exceeds expert human evaluation in chest X-ray interpretation (Figure 2). These results demonstrate substantial progress toward generalist medical AI systems capable of expert-level clinical reasoning across diverse diagnostic tasks.

Our comprehensive evaluation of multimodal LLMs for chest X-ray interpretation provides several key insights for medical ML applications:

- **Task-specific cognitive capabilities:** Our finding that the best model (MedGemma) achieves 85.03% accuracy on negation tasks but only 76.71% in differential diagnosis demonstrates that medical AI requires explicit design for different cognitive skills rather than treating all diagnostic reasoning as a uniform task.
- **Expert-guided dataset creation methodology:** Our three-layer pipeline with ra-

diologist validation offers a replicable approach for developing clinically representative datasets in other medical domains where direct annotation is costly or impractical.

- **Category-specific performance patterns:** Our detailed analysis across anatomical structures demonstrates that architectural decisions significantly impact performance on specific medical findings (e.g., while Janus-Pro-7B shows competitive performance on some skeletal structures, MedGemma demonstrates superior performance across nearly all anatomical categories including 94.04% on bone assessment and 97.03% on heart findings), suggesting specialized architectures may be more effective than general-purpose approaches for clinical applications.

2. Related Work

Early efforts in medical visual question answering (VQA) laid important groundwork but were limited in scope and complexity. VQA-RAD[10] introduced just 3,515 questions over 315 images, focusing primarily on basic anatomical queries. Similarly, ImageCLEF VQA-Med[1] offered binary questions like *"Is there something wrong in the image?"* suitable for feasibility studies but inadequate for training generalist systems. More recent datasets have expanded the scale and sophistication of medical VQA. PMC-VQA [19] introduced 227K question-answer pairs with free-text answer based on 149K diverse medical images from PubMed papers. MIMIC-CXR-VQA[5] provided 377K questions derived from radiology reports, but relied heavily on templated generation. Medical-Diff-VQA[7] took a novel approach by focusing on temporal reasoning over paired images, generating 700K questions for comparative assessment. MIMIC-Ext-MIMIC-CXR-VQA[3] improved the linguistic variety with paraphrased templates, while GEMeX[11] offered 1.6M multimodal questions with explanations. Despite these advances, most current datasets depend on rigid templates and fail to capture the flexible, multistep reasoning processes typical in radiology. Our work builds on this foundation but takes a distinct approach constructing questions via expert-refined prompts validated by radiologists to better reflect real clinical reasoning patterns and assess diverse cognitive capabilities.

3. The ReXVQA Dataset

In this section, we present the properties of ReXVQA dataset, a comprehensive multimodal benchmark for evaluating LLMs in radiology. We selected the MCQ format for its significant advantages over long-form assessment methodologies, as detailed in Table A.1 supplementary material. We discuss the data collection methodology, the preparation process, and the resulting dataset characteristics. A detailed visualization of the multi-stage MCQ generation pipeline used to create the ReXVQA dataset is provided in the supplementary material

3.1. *Task Definition*

The ReXVQA task can be formalized as $\mathbf{X}_i = (\mathbf{I}_i, \mathbf{Q}_i, \mathbf{O}_i)$, where $\mathbf{I}_i$ represents the i-th X-ray image input, $\mathbf{Q}_i$ represents the i-th question text, and $\mathbf{O}_i$ represents the set of candidate options. For each question-image pair, multiple candidate answers are provided as $\mathbf{O}_i = \{\mathbf{O}_{i1}, \mathbf{O}_{i2}, \mathbf{O}_{i3}, \mathbf{O}_{i4}\}$. The task requires models to analyze both the visual input $\mathbf{I}_i$ and the textual question $\mathbf{Q}_i$ to select the correct answer(s) from the option set. The ground truth label for each data point is defined as $y \in \mathbb{R}^1$ where $y^i = \{\mathbf{0}, \mathbf{1}, \mathbf{2}, \mathbf{3}\}$. The objective is to learn a prediction function $f : (I, Q) \rightarrow y$ that can effectively combine visual and textual information to make accurate diagnostic and clinical judgments. In addition, models are required to provide explanations $\mathbf{E}$ for their choices, making the complete prediction tuple $(\mathbf{y}, \mathbf{E})$. This explanation component allows for evaluation of the model's reasoning process and clinical understanding beyond mere answer selection.

MCQ Format Justification. We adopted the four-option MCQ format for several reasons. It mirrors established practices in medical education and board exams (e.g., USMLE, radiology boards), providing familiarity and clinical relevance. MCQs also enable systematic assessment of cognitive skills with objective, reproducible scoring, while the four-option design balances complexity and cognitive load.

Table A.1 in the supplementary material details the advantages of MCQs over long-form assessments, including scalability, consistent scoring, and precise analysis of reasoning patterns. While long-form responses capture nuanced clinical reasoning, they face challenges in standardized evaluation and cross-model comparison. Our approach focuses on systematically evaluating core radiological reasoning, with expert-validated questions ensuring clinical authenticity.

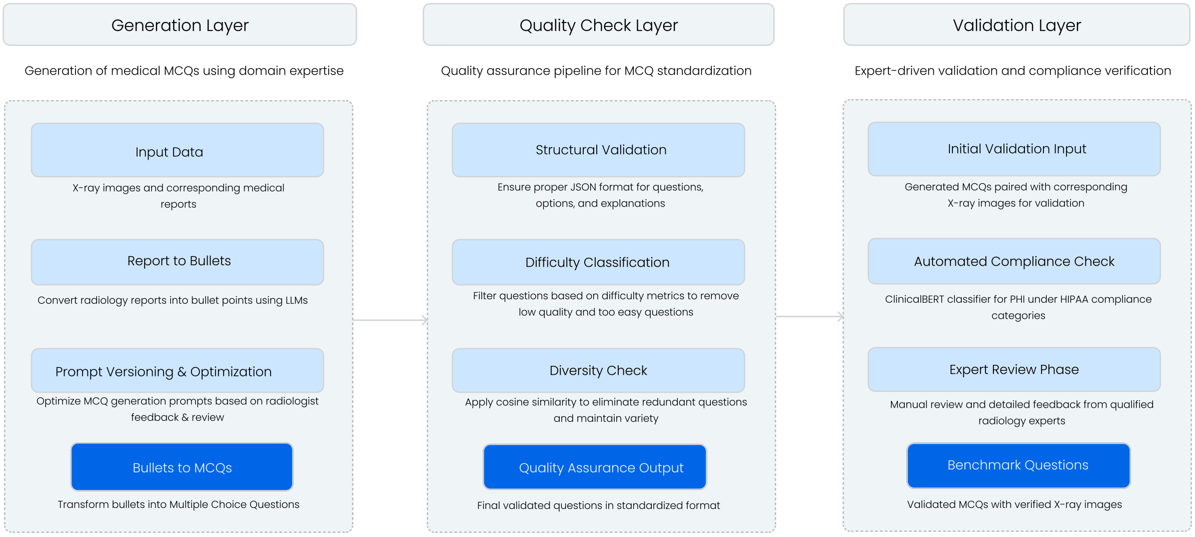
Generation Layer
Generation of medical MCQs using domain expertise
Input Data
X-ray images and corresponding medical reports
Report to Bullets
Convert radiology reports into bullet points using LLMs
Prompt Versioning & Optimization
Optimize MCQ generation prompts based on radiologist feedback & review
Bullets to MCQs
Transform bullets into Multiple Choice Questions
Quality Check Layer
Quality assurance pipeline for MCQ standardization
Structural Validation
Ensure proper JSON format for questions, options, and explanations
Difficulty Classification
Filter questions based on difficulty metrics to remove low quality and too easy questions
Diversity Check
Apply cosine similarity to eliminate redundant questions and maintain variety
Quality Assurance Output
Final validated questions in standardized format
Validation Layer
Expert-driven validation and compliance verification
Initial Validation Input
Generated MCQs paired with corresponding X-ray images for validation
Automated Compliance Check
ClinicalBERT classifier for PHI under HIPAA compliance categories
Expert Review Phase
Manual review and detailed feedback from qualified radiology experts
Benchmark Questions
Validated MCQs with verified X-ray images

Fig. 3. **Expert-Guided Medical MCQ Generation Pipeline:** We propose a three-layer approach combining computational processes and expert oversight for creating high-quality radiology MCQs.

3.2. *Source Dataset*

The source dataset, ReXGradient-160K,[17] comprises 170,000 chest X-ray studies with paired radiological reports from 109,722 unique patients across 4 U.S. health systems. This dataset represents the largest publicly available chest X-ray dataset to date in terms of patient count. The dataset is divided into public training (140,000 studies), public validation (10,000 studies), and public test (10,000 studies) sets, with an additional private test set, (10,000 studies).

Equipment Diversity The dataset encompasses chest X-rays acquired using equipment from multiple manufacturers including SIEMENS, FUJI, SAMSUNG, VIDAR, TOSHIBA, and GE. These manufacturers are distributed across all four hospital systems rather than being system-specific, reflecting real-world clinical diversity and enhancing model generalizability across different imaging technologies and acquisition protocols.

Expert Review Phase. Our expert review process incorporates a quality assurance protocol involving a board-certified radiologist. In total, 520 multiple-choice questions (MCQs) were reviewed. 300 were qualitatively assessed during initial development process to examine the LLM's question-answering quality, and 220 were evaluated quantitatively with error classification after refining prompts and models. These 220 questions spanned varying difficulty levels and included a dedicated image alignment assessment. The evaluation emphasized four critical dimensions: clinical quality, explanation clarity, factual correctness, and image alignment. The image alignment analysis revealed only one case (0.5%) of content-radiograph misalignment due to a source data discrepancy, suggesting minimal inconsistency resulting from our report-based approach. Figure A.1 in the supplementary material illustrates the radiology image tagging platform used for expert annotation.

Expert Review Outcomes. The quantitative expert review of 220 sample questions revealed several issues requiring correction. Specifically, 10.8% of questions had multiple valid answers, 6.7% contained unnecessary comparison-related content, 5% required improvements to ensure clinical validity, and 3.3% included hallucinated information about findings not described in the radiology reports. The most common issue occurred with negation-type questions, where multiple valid answers arose for findings that were absent in the original reports. To address these concerns, we implemented multiple validation steps, including careful cross-checking of the original reports and generated questions to remove comparison-related content and eliminate questions with multiple valid answers. This feedback directly informed our prompt engineering iterations, leading to refined question generation strategies, simplified medical language, standardized anatomical terminology, and ultimately proving highly effective, Only two errors were identified in a subsequent reader study involving 300 questions.

Benchmark Question Finalization. After rigorous multi-stage validation and quality control, questions meeting all quality criteria are incorporated into the final ReXVQA benchmark, facilitating detailed analysis of model performance across different dimensions of radiological expertise.

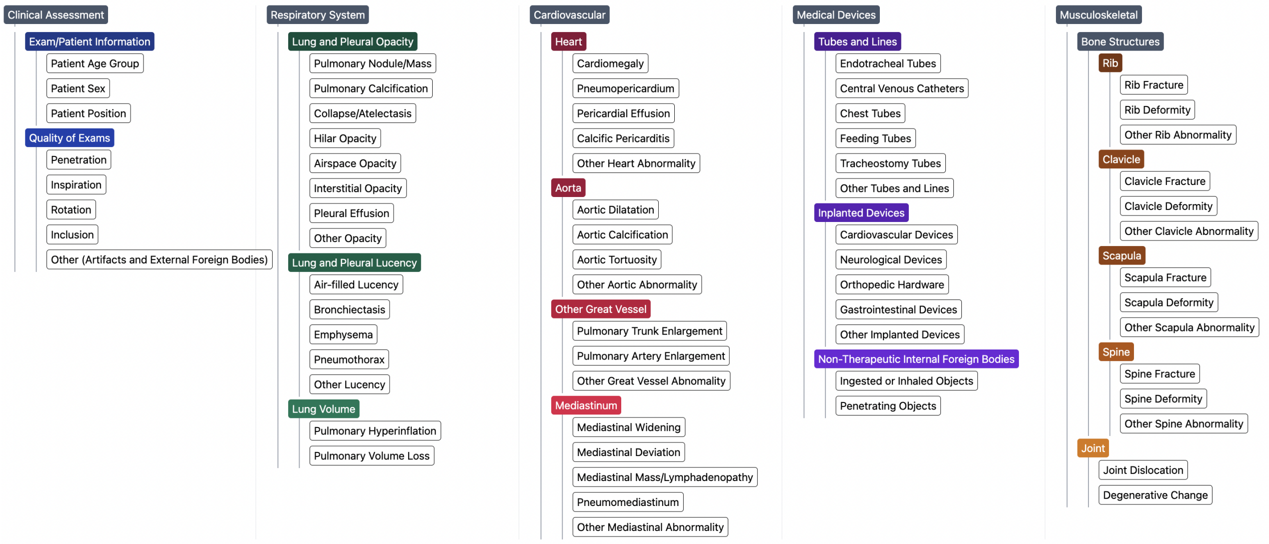

Fig. 4. **Hierarchical Taxonomy of Chest X-Ray Categories.** This expert-validated classification system, developed in collaboration with radiologists, organizes chest X-ray findings into five major domains: Clinical Assessment, Respiratory System, Cardiovascular, Medical Devices, and Musculoskeletal findings. The taxonomy serves as a structured foundation for Expert-Guided Medical MCQ generation, ensuring comprehensive coverage and clinical relevance.

3.3. *Cognitive Framework Development*

The five cognitive abilities evaluated in ReXVQA were informed by established question types in medical VQA literature and clinical practice patterns. Presence and negation assessment reflect the most common question types in VQA-RAD[9] and MIMIC-Ext-MIMIC-CXR-VQA,[4] with negation detection being clinically critical.[6] Location assessment aligns with anatomy-focused questions in prior datasets,[4,8] while differential diagnosis corresponds to abnormality detection tasks in VQA-Med.[2] Geometric analysis addresses quantitative measurements relevant to clinical assessment.[4]

Through radiologist consultation during prompt development, we validated these abilities as reflecting core radiological reasoning patterns, with task distribution prioritizing fundamental skills (∼73% for presence/negation) while ensuring representation of specialized assessment capabilities.

3.4. *Dataset Statistics*

For ReXVQA, we follow the same split as the source dataset. The private test set is reserved for independent evaluation through our leaderboard system, ensuring unbiased assessment of model performance. The dataset consists of 572,419 VQA pairs for training, 40,878 for validation, 40,826 for public testing, and 41,007 for private testing. The ReXVQA dataset encompasses a diverse range of radiological aspects, carefully structured to evaluate different dimensions of multimodal LLM capabilities in medical imaging interpretation.

3.4.1. *Task Distribution Analysis*

Our dataset incorporates five distinct task types across training, validation, test, and private test sets, with remarkably consistent distributions. As shown in Table 1, Negation Assessment and Presence Assessment together comprise approximately 72% of all tasks, highlighting their fundamental importance in radiological interpretation. Differential Diagnosis represents about 21% of tasks, while Location and Distribution Assessment (approximately 6%) and Geometric Information Assessment (less than 0.5%) target more specialized interpretative skills. This distribution reflects the hierarchical nature of radiological reasoning, from basic detection to complex spatial and differential analysis.

Table 1. Distribution of task categories across datasets. Private test set (41,007 cases) is reserved for leaderboard evaluation and not included in the publicly available dataset.

Category	**Train**	**Valid**	**Test**	**Private**
	Count (%)	Count (%)	Count (%)	Count (%)
Task Categories				
Negation Assessment	209,053 (**36.5**)	15,007 (**36.7**)	15,369 (**37.9**)	15,408 (**37.6**)
Presence Assessment	206,880 (**36.1**)	14,698 (**36.0**)	14,078 (**34.7**)	14,452 (**35.2**)
Differential Diagnosis	119,111 (**20.9**)	8,578 (**21.0**)	8,563 (**21.1**)	8,585 (**20.9**)
Location & Distribution	34,829 (**6.1**)	2,404 (**5.9**)	2,365 (**5.8**)	2,383 (**5.8**)
Geometric Information	2,546 (**0.4**)	171 (**0.4**)	182 (**0.5**)	179 (**0.4**)
Total	572,419	40,858	40,557	41,007

3.4.2. *Anatomical Category Distribution*

Analysis across the dataset reveals a diverse but clinically realistic distribution of anatomical categories. Lung and Pleural Opacity dominates (30.2-30.4%), reflecting the prevalence of this finding in chest radiography, followed by Heart assessments (14.6-15.0%) and Negation (13.2-13.5%). The distribution encompasses supportive devices such as Tubes and Lines (5.0-5.2%), along with Other Pulmonary Diagnosis (4.2-4.5%). The dataset maintains balanced representation across critical diagnostic areas, including Infectious Disease (2.3-2.5%), Pulmonary Vascularity (1.8%), and Cardiac Disease (0.35%), while also covering essential supporting structures such as Spine, Ribs, and Mediastinum. Importantly, the distribution captures both common conditions and rare but clinically significant findings like Pulmonary Neoplasm (0.32-0.35%) and Lymphoproliferative Disease (0.02-0.03%), along with technical quality assessments (3.4-3.6%). This distribution mirrors real-world clinical prevalence while ensuring sufficient representation for comprehensive model evaluation. Table A.2 in the supplementary material presents the taxonomy of medical conditions in our dataset, categorizing them into nine main classes with their respective subcategories.

4. Experiments

4.1. *Baseline Models*

The primary objective of our baseline experiments is to evaluate the performance of current state-of-the-art multimodal LLMs on ReXVQA, specifically focusing on their ability to handle complex radiological MCQs designed for medical professionals. We selected models with varying architectures, training approaches, and accessibility to provide a comprehensive benchmark. Our evaluation includes both commercial and open-source models, representing the current landscape of multimodal AI capabilities in medical imaging. Notably, MedGemma represents a medical-domain-specific model, allowing us to compare general-purpose multimodal models against specialized medical AI systems.

For brevity, we refer to the evaluated models using the following short names throughout the paper: Phi35 (Phi-3.5-vision-instruct), Qwen2VL (Qwen2-VL), Qwen25VL (Qwen2.5-VL), Gemini (Gemini 1.5 Pro), Eagle2, Janus, LLaVA, and MedGemma. A detailed description of each model, including architecture and training background, is provided in the supplementary material. Importantly, none of the evaluated models overlap with the LLM used for dataset generation (GPT-4o), ensuring unbiased evaluation without data leakage or model-specific advantages.

4.2. *Evaluation Framework*

Our evaluation framework implements a standardized protocol for assessing model performance. For each query, models receive an X-ray image, accompanied by a question in natural language and four multiple-choice options. Models must provide their selected option and a detailed explanation justifying their choice for their prediction. We employ the standard accuracy as the evaluation metric.

Image Input Specifications. Models receive chest X-ray images in PNG format (converted from original DICOM files). For the public dataset, images are provided at 1/4 of original resolution to balance computational efficiency with diagnostic detail preservation. This preprocessing maintains clinically relevant features while enabling scalable evaluation across multiple models. For studies containing multiple radiographic views (as occurs in real-world radiology practice), models are provided with all available images paired with each question, enabling comprehensive assessment across different anatomical projections.

5. Results and Analysis

The comprehensive evaluation of eight state-of-the-art multimodal models revealed significant variations in their ability to interpret chest X-rays across different clinical domains. MedGemma demonstrated exceptional performance, achieving 83.24% overall accuracy and establishing a new benchmark for multimodal medical image interpretation. This represents a substantial improvement over previous leading models, with Janus-Pro-7B following at 66.56%, followed closely by Qwen25VL (65.55%) and Eagle2 (64.43%) as shown in Table 2. Gemini achieves a respectable 63.31%, while LLaVA struggles notably with only 26.61% accuracy, highlighting the considerable challenges in multimodal medical image interpretation.

Table 2. Evaluation of models on various diagnostic and assessment metrics. Failed Extractions shows the percentage of test cases where models failed to provide valid responses in the required format (out of 41,007 private test cases). ↑ indicates higher is better, ↓ indicates lower is better.

Model	Overall Accuracy ↑	Differential Diagnosis ↑	Geometric Information ↑	Location and Distribution ↑	Negation Assessment ↑	Presence Assessment ↑	Failed Extractions ↓
LLaVA	26.61	21.61	23.46	27.61	24.02	36.33	2.37
Phi35	47.49	62.24	22.15	37.11	79.50	36.44	0.05
Qwen2VL	54.70	52.65	44.94	54.05	62.69	59.15	0.01
Gemini	63.31	62.21	46.89	59.60	85.68	62.17	**0.0**
Eagle2	64.43	68.17	56.98	56.95	**86.32**	53.75	**0.0**
Qwen25VL	65.55	63.61	66.48	63.24	83.27	51.14	**0.0**
Janus-Pro-7B	66.56	56.34	75.42	64.62	75.73	60.70	**0.0**
MedGemma	**83.24**	**76.71**	**80.45**	**83.47**	85.03	**85.21**	**0.0**

5.1. *Task-Specific Performance Analysis*

The models demonstrate distinct strengths across different radiological reasoning tasks, with MedGemma leading in four out of five major categories as shown in Table 2. Based on our analysis of model architectures and performance patterns:

- **Differential Diagnosis**: MedGemma achieves the highest performance at 76.71%, substantially outperforming second model Eagle2 with accuracy 68.17%. This superior performance suggests MedGemma's specialized medical training enables more sophisticated clinical reasoning for distinguishing between similar conditions.
- **Geometric Information Assessment**: MedGemma excels with 80.45% accuracy, surpassing Janus-Pro-7B's 75.42%. This improvement indicates enhanced capabilities for spatial representation and precise measurement interpretation in radiological contexts.
- **Location and Distribution Assessment**: MedGemma leads significantly at 83.47%, well above Janus-Pro-7B's 64.62% and Qwen25VL's 63.24%. This performance suggests superior positional representation mechanisms for localizing findings within complex radiological images.
- **Negation Assessment**: Eagle2 achieves the best performance at 86.32%, closely followed by Gemini (85.68%) and MedGemma (85.03%). The top three models demonstrate consistently high standards for identifying absence of findings a critical skill in avoiding false positives.
- **Presence Assessment**: MedGemma demonstrates exceptional capability at 85.21%, substantially exceeding Gemini's previous best of 62.17%. This dramatic improvement suggests superior feature extraction capabilities for detecting radiological abnormalities within complex backgrounds.

5.2. *Category-wise Performance Analysis*

Table 3 & Table A.3 in the supplementary material presents a detailed breakdown of model performance across key radiological categories, with MedGemma consistently outperforming other models across most categories, achieving an average performance of 83.24%.

Table 3. Comprehensive performance comparison of models across key radiological categories (values shown in %). Bold numbers indicate best performance per category. Abbreviations: P.O. = Pleural Opacity, P.L. = Pleural Lucency and Q2 & Q25 = Qwen2 & Qwen25

Category	**Gemini**	**Eagle2**	**Janus**	**LLaVA**	**Q2VL**	**Q25VL**	**Phi35**	**MedGemma**
Clinical Assessment								
Quality of Exams	70.30	62.50	51.64	10.30	40.50	60.92	35.10	**71.23**
Respiratory System								
Lung & P.O	72.24	72.19	68.70	32.98	60.73	64.77	58.14	**80.44**
Lung & P.L	80.00	64.65	58.59	19.34	61.62	78.64	58.82	**87.88**
Lung Volume	65.44	58.22	51.64	39.72	52.11	60.92	28.25	**78.64**
Cardiovascular								
Heart	80.29	81.71	72.01	26.01	60.31	84.83	62.73	**97.03**
Aorta	76.65	41.04	60.14	6.84	72.17	34.79	33.05	**87.86**
Other Great Vessel	**73.33**	60.00	60.00	13.33	60.00	60.00	45.45	**73.33**
Medical Devices								
Tubes and Lines	59.45	58.26	58.87	22.81	48.78	65.04	32.10	**83.86**
Implanted Devices	54.46	52.64	60.79	29.50	52.40	53.48	43.16	**73.14**
Pathologies								
Infectious Disease	71.55	63.24	66.77	56.63	51.02	67.74	42.13	**77.61**
Pulmonary Neoplasm	78.46	66.92	81.95	39.29	63.91	73.68	69.44	**88.72**
Negation	61.05	71.73	58.19	15.00	56.37	61.49	**78.96**	74.76
Musculoskeletal								
Rib	88.90	83.93	89.80	36.92	86.35	79.21	78.56	**91.84**
Spine	78.84	62.30	86.43	64.84	73.73	50.10	57.94	**92.68**
Clavicle	75.93	57.14	75.00	23.21	71.43	33.93	51.02	**92.73**
Joint	51.96	42.31	70.19	47.06	66.35	36.54	42.22	**88.46**
Average	74.00	67.00	67.00	38.00	64.00	66.00	50.00	**83.24**

Clinical Assessment. In exam quality interpretation, MedGemma demonstrates superior capabilities (71.23%), followed by Gemini (70.30%), Eagle2 (62.50%) and Qwen25VL (60.92%). LLaVA's performance (10.30%) suggests significant limitations in understanding technical image characteristics. This disparity indicates that advanced multimodal architectures are essential for capturing the nuanced details required for technical quality assessment.

Respiratory System. Respiratory findings analysis reveals consistent performance patterns across subcategories. MedGemma leads in all three respiratory metrics, achieving 80.44% for lung and pleural opacities, 87.88% for pleural lucencies, and 78.64% for lung volume assessment. The substantial performance gap between top models and LLaVA (19.34-39.72%) underscores the complexity of pulmonary pattern recognition.

Cardiovascular Imaging. Cardiovascular interpretation presents interesting variations across subcategories. MedGemma leads in heart finding analysis (97.03%), substantially exceeding Qwen25VL's 84.83%. and 87.86% for aortic assessment (compared to Gemini's

76.65%). This consistent excellence across different vascular structures suggests robust architectural capabilities for cardiovascular imaging, addressing the previous inconsistencies observed among other models.

Medical Devices Detection. MedGemma significantly advances medical device recognition with 83.86% for tubes and lines detection and 73.14% for implanted devices. These improvements suggest that specialized medical training helps models better understand artificial structures despite their variable appearance and positioning.

Pathologies. MedGemma demonstrates superior pathology identification capabilities: 88.72% for pulmonary neoplasms, 77.61% for infectious diseases, and 74.76% for negation assessment (Phi35 maintains a slight edge at 78.96%). These results confirm that medical domain specialization enhances pattern recognition for diverse pathological conditions.

Musculoskeletal Findings. MedGemma achieves exceptional performance in skeletal structure assessment, leading all categories: 91.84% for rib interpretation, 92.68% for spine assessment, 92.73% for clavicle detection, and 88.46% for joint interpretation. These results demonstrate that even for high-contrast bony structures that were already well-recognized by previous models, specialized medical training can yield substantial improvements.

6. Reader Studies

6.1. *Overall Performance Analysis*

Our reader study evaluated the diagnostic performance of AI models compared to three radiology residents in their 3rd or 4th year of training, using 200 randomly sampled chest X-ray cases. The results reveal that current AI models can achieve competitive performance with human readers in standard diagnostic tasks, although there are significant variations between different models. Among the AI models tested, MedGemma demonstrated the highest overall accuracy at 83.84%, substantially outperforming all other models and human readers. Qwen25VL achieved 77.78% accuracy, closely matching the performance of the top human reader (Reader 3 at 77.27%). Reader 2 achieved 69.70% accuracy, while Reader 1 performed at 65.66%, comparable to several AI models, including Janus (66.16%) and Eagle2 (64.14%). The performance distribution shows a clear hierarchy, with some models like LLaVA (24.75%) and Phi35 (37.88%) demonstrating significantly lower accuracy, indicating substantial variability in current AI model capabilities for medical image interpretation.

This variability is further reflected in the interrater agreement patterns across models and human readers (see Section A.4 and Figure A.3 in the supplementary material)

Acknowledgement

This work was supported by the Biswas Family Foundation's Transformative Computational Biology Grant in Collaboration with the Milken Institute.

References

1. Asma Ben Abacha, Vivek Datla, Sadid A. Hasan, Dina Demner-Fushman, and Henning Müller. Overview of the vqa-med task at imageclef 2020: Visual question answering and generation in the medical domain. In *Conference and Labs of the Evaluation Forum*, 2020.
2. Asma Ben Abacha, Sadid A. Hasan, Vivek Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Conference and Labs of the Evaluation Forum*, 2019.
3. Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, and Edward Choi. Mimic-ext-mimic-cxr-vqa: A complex, diverse, and large-scale visual question answering dataset for chest x-ray images, 2024.
4. Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, and Edward Choi. MIMIC-Ext-MIMIC-CXR-VQA: A Complex, Diverse, And Large-Scale Visual Question Answering Dataset for Chest X-ray Images (version 1.0.0). `https://doi.org/10.13026/deqx-d943`, July 2024. PhysioNet. RRID:SCR_007345.
5. Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, et al. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Advances in Neural Information Processing Systems*, 36, 2024.
6. Sergey Goryachev, Margarita Sordo, Qing T. Zeng, and Long H. Ngo. Implementation and evaluation of four different methods of negation detection. 2007.
7. Xinyue Hu, Lin Gu, Qi A. An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M. Summers, and Yingying Zhu. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.
8. Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, liangchen liu, Kazuma Kobayashi, Tatsuya Harada, Ronald Summers, and Yingying Zhu. Medical-Diff-VQA: A Large-Scale Medical Dataset for Difference Visual Question Answering on Chest X-Ray Images (version 1.0.1). `https://doi.org/10.13026/e6dd-cn74`, Feb 2025. PhysioNet. RRID:SCR_007345.
9. Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. 2018.
10. Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. Descriptor : A dataset of clinically generated visual questions and answers about radiology images. 2018.
11. Bo Liu, Kevin Yingyin Zou, Li-Ming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. *ArXiv*, abs/2411.16778, 2024.
12. Michael Moor, Oishi Banerjee, Zahra F H Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616:259–265, 2023.
13. Ankit Pal and Malaikannan Sankarasubbu. Gemini goes to med school: Exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. In *Clinical Natural Language Processing Workshop*, 2024.
14. Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12):1399–1406, 2022.
15. Sheng Wang, Zihao Zhao, Ouyang Xi, Tianming Liu, Qian Wang, and Dinggang Shen. Interactive computer-aided diagnosis on medical image using large language models. *Communications Engineering*, 3, 2024.

16. Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21372–21383, 2023.
17. Xiaoman Zhang, Julián N. Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. Rexgradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports. In *arXiv:2505.00228v1*, 2025.
18. Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542, 2023.
19. Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4(1):277, 2024.