Appendix A.

1. Baseline Models Details

- **LLaVA 1.5** (2023): An open-source chatbot model trained by fine-tuning LLaMA/Vicuna on GPT-generated multimodal instruction-following data. The model connects a CLIP ViT-L/14 visual encoder with Vicuna using a projection matrix and was trained on 158K unique language-image instruction samples.[10]
- **Phi-3.5-vision-instruct** (Microsoft, 2024): A lightweight multimodal model (4.2B parameters) supporting 128K token context length. Trained on 500B tokens across vision and text data, it excels at multi-frame understanding and image comparison tasks.[9]
- **Qwen2-VL** (Alibaba, 2024): An open-source vision-language model with dynamic resolution capabilities that maintains original aspect ratios without distortion.[1]
- **Qwen2.5-VL** (Alibaba, 2025): An upgraded model with enhanced recognition of handwritten text and multiple languages. It introduces structured output capabilities for data extraction and accurate object localization through bounding boxes or points.[2]
- **Gemini 1.5 Pro** (Google, 2024): A multimodal model designed for complex reasoning with extensive context processing capabilities. It achieves near-perfect recall on long-context retrieval tasks and demonstrates significant improvements in document and video question answering.[4]
- **Eagle2-9B** (NVIDIA, 2024): A vision-language model balancing performance and inference speed. It combines SigLip and ConvNext vision encoders with Qwen2.5-7B-Instruct.[6]
- **Janus-Pro-7B** (DeepSeek, 2025): An autoregressive framework unifying multimodal understanding and generation. Built on DeepSeek-LLM-7b-base with SigLIP-L vision encoder, it supports 384×384 image inputs and significantly outperforms its predecessor on multimodal understanding benchmarks.[3]
- **MedGemma** (Google, 2025): MedGemma 4B utilizes a SigLIP image encoder that has been specifically pre-trained on a variety of de-identified medical data, including chest X-rays, dermatology images, ophthalmology images, and histopathology slides. Its LLM component is trained on a diverse set of medical data, including radiology images, histopathology patches, ophthalmology images, and dermatology images.[5]

1.1. *Dataset Creation Pipeline*

ReXVQA proposed a three-layer pipeline architecture designed to transform raw radiological data into high-quality MCQs while maintaining clinical accuracy and educational value. Importantly, our pipeline utilizes radiology reports exclusively to avoid potential multimodal hallucination and ensure clinical accuracy. Figure A2 presents the architectural overview.

1.1.1. *Generation Layer*

Input Data Processing. The foundation of our pipeline consists of paired X-ray images and their corresponding medical reports from our curated dataset. Each report undergoes

Table A1. Comparison of evaluation methodologies for radiological LLM assessment. + indicates advantage, − indicates limitation in the below table

Criteria	**Long-Form**	**MCQ**
Reproducibility	Limited −	High +
Standardization	Variable −	Consistent +
Quantification	Subjective −	Objective +
Resources	High −	Low +
Scalability	Limited −	Extensive +
Granularity	Coarse −	Fine +
Reasoning	Implicit −	Explicit +
Automation	Limited −	High +

Table A2. Classification of Medical Conditions and their Subcategories

Main Category	**Subcategories**
Congenital Disease	Congenital Lung Disease, Congenital Vascular Disease, Congenital Heart Disease
Infectious Disease	Pneumonia, Tuberculosis, Other Infection
Pulmonary Neoplasm	Primary Lung Cancer, Pulmonary Metastases, Other Pulmonary Neoplasm
Lymphoproliferative Disease	Lymphoma, Other Lymphoproliferative Disease
Other Pulmonary Diagnosis	Interstitial Lung Disease, Sarcoidosis, Asbestos-Related Disease, Pneumoconiosis, Pulmonary Edema, ARDS, Aspiration, Iatrogenic Lung Disease, COPD, Vasculitis, Pulmonary Hypertension, Pulmonary Thromboembolic Disease, Miscellaneous Pulmonary Disease
Cardiac Disease	Valvular Heart Disease, Myocardial Disease, Pericardial Disease, Congestive Heart Failure, Other Cardiac Disease
Aortic Disease	Aortic Dissection/Aneurysm, Atherosclerosis, Other Aortic Disease
Miscellaneous Diagnosis	Trauma, Post-Treatment Change, Miscellaneous Disease
Negation	Absence of Disease

extensive preprocessing through our medical-domain-specific pipeline.

Report to Bullets Transformation. We utilized GPT-4o (version 2024-05-01-preview, Azure OpenAI API) to systematically transform radiology reports into structured bullet points. This crucial step preserves the complete clinical context while presenting the information in a more organized format. The transformation process focuses on maintaining all critical findings, anatomical descriptions, and diagnostic interpretations from the original report. Our prompt engineering ensures that the bullet points capture both normal and abnormal findings, maintaining the hierarchical structure of radiological observations. The resulting bullet points

Category	Gemini	Eagle2	Janus-Pro-7B	LLaVA	Qwen2VL	Qwen25VL	Phi35	MedGemma
Abdomen	75.34	55.48	52.05	28.77	59.59	61.64	32.50	78.77
Airway	69.84	71.28	56.66	24.54	60.84	59.01	81.61	92.17
Aorta	76.65	41.04	60.14	6.84	72.17	34.79	33.05	87.86
Aortic Disease	45.00	28.57	19.05	9.52	14.29	19.05	25.00	33.33
Bone	87.99	89.82	86.67	31.93	91.93	88.77	80.41	94.04
Bone Density	50.00	50.00	0.00	100.00	100.00	0.00	0.00	50.00
Bones	100.00	100.00	100.00	25.00	100.00	91.67	87.50	100.00
Bones and Soft Tissues	100.00	100.00	100.00	100.00	100.00	100.00	-	100.00
Bones/Joints	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Bony Structures	100.00	100.00	100.00	60.00	100.00	100.00	100.00	100.00
Bony Thorax	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Cardiac Disease	59.03	80.69	28.97	30.56	33.10	95.17	23.62	80.56
Chest Wall	60.00	30.00	70.00	30.00	60.00	70.00	37.50	40.00
Clavicle	75.93	57.14	75.00	23.21	71.43	33.93	51.02	92.73
Congenital Disease	71.43	71.43	57.14	40.00	71.43	42.86	33.33	57.14
Diaphragm	64.16	51.18	40.83	8.61	61.83	42.60	44.24	68.93
Gastrointestinal Devices	16.67	33.33	66.67	50.00	33.33	16.67	25.00	60.00
Heart	80.29	81.71	72.01	26.01	60.31	84.83	62.73	97.03
Hila	100.00	100.00	0.00	0.00	0.00	100.00	100.00	100.00
Hilar Opacity	100.00	100.00	100.00	0.00	100.00	100.00	100.00	100.00
Humerus	67.35	56.86	82.35	39.22	66.67	47.06	52.38	84.00
Implanted Devices	54.46	52.64	60.79	29.50	52.40	53.48	43.16	73.14
Infectious Disease	71.55	63.24	66.77	56.63	51.02	67.74	42.13	77.61
Joint	51.96	42.31	70.19	47.06	66.35	36.54	42.22	88.46
Lung Volume	65.44	58.22	51.64	39.72	52.11	60.92	28.25	78.64
Lung and Pleural Lucency	80.00	64.65	58.59	19.34	61.62	78.64	58.82	87.88
Lung and Pleural Opacity	72.24	72.19	68.70	32.98	60.73	64.77	58.14	80.44
Lymph Nodes	100.00	11.11	0.00	0.00	0.00	100.00	0.00	22.22
Lymphoproliferative Disease	90.00	60.00	50.00	30.00	60.00	80.00	57.14	60.00
Mediastinum	93.27	79.77	85.93	15.24	70.85	90.83	80.73	90.58
Miscellaneous Bone Abnormality	52.43	50.49	62.62	39.32	65.85	43.20	40.74	74.15
Miscellaneous Diagnosis	68.83	64.94	72.73	31.51	61.04	68.83	55.17	79.22
Musculoskeletal	0.00	0.00	100.00	100.00	0.00	0.00	0.00	100.00
Neck/Chest Wall Soft Tissue	85.94	77.27	56.06	27.27	66.67	51.52	41.38	83.33
Negation	61.05	71.73	58.19	15.00	56.37	61.49	78.96	74.76
Non-Therapeutic Internal Foreign Bodies	67.86	63.33	65.00	26.67	65.00	66.67	53.85	74.58
Osseous Structures	95.24	100.00	100.00	23.81	100.00	100.00	100.00	100.00
Osseous structures and Soft Tissues	100.00	100.00	100.00	0.00	100.00	100.00	100.00	100.00
Other (Artifacts and External Foreign Bodies)	50.00	50.00	0.00	0.00	0.00	50.00	0.00	100.00
Other Bone Abnormality	0.00	0.00	0.00	100.00	0.00	100.00	0.00	100.00
Other Great Vessel	73.33	60.00	60.00	13.33	60.00	60.00	45.45	73.33
Other Implanted Devices	100.00	100.00	0.00	0.00	0.00	100.00	0.00	100.00
Other Pulmonary Diagnosis	59.85	59.85	45.51	22.07	43.00	64.77	31.14	81.26
Pleura	100.00	100.00	100.00	0.00	33.33	100.00	100.00	100.00
Pleural	100.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
Pleural Effusion	87.88	93.94	42.42	6.06	39.39	84.85	75.00	69.70
Pleural Space	100.00	100.00	100.00	0.00	100.00	100.00	0.00	100.00
Pleural Thickening	79.84	75.78	67.97	23.44	58.59	71.88	57.43	82.81
Pulmonary Fissure	90.00	60.00	80.00	40.00	60.00	30.00	75.00	90.00
Pulmonary Neoplasm	78.46	66.92	81.95	39.29	63.91	73.68	69.44	88.72
Pulmonary Vascularity	73.30	67.88	82.39	25.54	38.31	78.63	50.44	83.31
Rib	88.90	83.93	89.80	36.92	86.35	79.21	78.56	91.84
Scapula	28.57	14.29	42.86	28.57	28.57	14.29	33.33	57.14
Shoulder	75.00	50.00	50.00	50.00	100.00	50.00	0.00	100.00
Skeletal Structures	98.86	99.62	99.87	32.83	100.00	99.87	99.82	100.00
Skeletal System	70.00	90.00	80.00	80.00	90.00	70.00	87.50	100.00
Skeleton	100.00	100.00	100.00	75.00	100.00	100.00	100.00	100.00
Soft Tissue	60.00	80.00	80.00	20.00	60.00	80.00	25.00	100.00
Spine	78.84	62.30	86.43	64.84	73.73	50.10	57.94	92.68
Sternum	92.50	85.37	87.80	70.00	87.80	87.80	75.00	92.68
Trauma	60.00	60.00	100.00	60.00	60.00	80.00	80.00	60.00
Tubes and Lines	59.45	58.26	58.87	22.81	48.78	65.04	32.10	83.86
Vascular	0.00	100.00	100.00	100.00	100.00	0.00	0.00	100.00
Vasculature	100.00	100.00	100.00	100.00	100.00	100.00	0.00	100.00
Vascularity	-	0.00	100.00	100.00	100.00	0.00	0.00	100.00
Average	74.00	67.00	67.00	38.00	64.00	66.00	50.00	83.24

serve as an intermediate representation that bridges the gap between unstructured reports and structured MCQ generation while ensuring no critical information is lost in the process.

Prompt Versioning and Optimization. Our prompt engineering process underwent twelve major iterations to optimize the quality and clinical accuracy of generated MCQs. Each iteration was refined through systematic feedback from a board-certified radiologist who evaluated the generated questions based on three key criteria: question quality, clinical accuracy, and educational value. The prompt templates were iteratively improved to address specific challenges identified during the review process, such as ensuring questions test interpretive skills rather than mere recall, incorporating appropriate distractors, and maintaining clinical relevance.

MCQ Generation. The final step employs our specialized medical prompt template with GPT-4o (version 2024-05-01-preview, Azure OpenAI API) to transform bullet points into

MCQs.

1.1.2. *Quality Check Layer*

We implement a comprehensive validation framework that enforces strict structural and content requirements through two distinct validator types: structural validators and content validators.

Structural Validators (v_i): These validators ensure JSON schema compliance and data format integrity:

- **Schema Compliance Validator (v_1):** Verifies that each MCQ conforms to the required JSON structure with mandatory fields.
- **Data Type Validator (v_2):** Ensures all fields contain appropriate data types (strings for text, integers for indices, structured objects for metadata).
- **Format Validator (v_3):** Checks that answer options are properly formatted, explanations meet minimum length requirements, and metadata contains required difficulty and category classifications.
- **Completeness Validator (v_4):** Verifies no required fields are empty or null.

Content Validators (c_j): These validators assess medical accuracy and educational value:

- **Domain Specificity Validator (c_1):** Questions should test radiology knowledge and not be answerable via general internet searches.
- **Cognitive Depth Validator (c_2):** Questions must require interpretative reasoning rather than simple recall.
- **Clinical Alignment Validator (c_3):** Answers and explanations must align with clinical guidelines and best practices.

Each MCQ must conform to our JSON schema:

$$Q = \{q, \{o_1, o_2, o_3, o_4\}, a, e, m\},$$

where q represents the question text, o_i are answer options, a is the correct answer index, e is the detailed explanation, and m contains metadata including difficulty level and clinical categories. Our validation system employs a logical AND-based checking mechanism:

$$V(Q) = (v_1(Q) \wedge v_2(Q) \wedge ... \wedge v_k(Q)) \wedge (c_1(Q) \wedge ... \wedge c_l(Q)), \tag{A.1}$$

where $k = 4$ structural validators and $l = 3$ content validators. Each validator outputs a binary value (0 or 1), and $V(Q)$ is true if and only if all validators return true.

Difficulty Classification. We implemented a systematic approach to difficulty calibration and quality control in our MCQ dataset. The classification process operates at two levels: automated LLM-based assessment and expert validation. During question generation, we instructed the LLM to provide an initial difficulty rating for each question, categorizing them into three tiers: easy, medium, and hard:

$$D(Q) = \text{LLM}(Q, C_{diff}), \tag{A.2}$$

where $D(Q)$ represents the difficulty score and C_{diff} encompasses our difficulty criteria framework. To validate this automated classification, we employed stratified random sampling to ensure coverage across all difficulty levels. Specifically, for every 1,000 generated MCQs, we sampled 100 questions evenly across the three difficulty tiers. This process ensures that questions of varying complexity are proportionally represented during expert review. Our validation process specifically focused on ensuring questions met two critical criteria:

- **Domain Specificity**: Questions must require radiological expertise and cannot easily be answered through simple internet searches.
- **Cognitive Depth**: Each question should test interpretative skills and clinical reasoning rather than mere fact recall

Questions found to be either too elementary or lacking in radiological specificity were filtered out using our quality threshold. This rigorous filtering process ensures that our benchmark maintains appropriate difficulty levels for evaluating advanced radiological knowledge and reasoning capabilities.

Diversity Check. To ensure comprehensive coverage of radiological concepts while avoiding redundancy in our benchmark, we implemented a diversity assessment system. This system evaluates the similarity between question pairs using both semantic and structural features. For semantic similarity, we leverage MedEmbed embeddings to capture the underlying meaning and clinical concepts in each question.[8]

$$Div(Q_i, Q_j) = \lambda_1 sim_{text}(Q_i, Q_j), \tag{A.3}$$

where sim_{text} measures semantic similarity using MedEmbed embeddings. Questions are filtered if:

$$Div(Q_i, Q_j) > \tau_{diversity}, \tag{A.4}$$

where $\tau_{diversity}$ is empirically set to 0.9 based on expert evaluation. During our validation process, a radiologist reviewed pairs of questions with varying similarity scores to establish this optimal threshold.

1.1.3. *Validation Layer*

Initial Validation Input. The validation process begins with a structured input tuple:

$$V_{in} = (Q_i, I_i, M_i, H_i)_{i=1}^{n}, \tag{A.5}$$

where Q_i represents the MCQ, I_i the X-ray image, M_i the metadata, and H_i the generation history. Each component undergoes independent validation before proceeding to compliance checking.

Automated Compliance Check. We employ ClinicalBERT,[7] a BERT-variant specifically fine-tuned classifier on medical compliance data. This system performs multi-faceted compliance checking across three critical dimensions: Protected Health Information (PHI) detection,

HIPAA compliance verification, and bias assessment. The system performs sequential checks.[11] The compliance classifier outputs a binary decision for each question:

$$C(Q) = \begin{cases} 1 & \text{if } P_{phi} \geq \theta_c \\ 0 & \text{otherwise,} \end{cases} \tag{A.6}$$

where θ_c is our compliance threshold. Questions failing any compliance check are automatically filtered from the dataset. This rigorous screening process ensures that our benchmark maintains high standards of privacy protection and fairness while preserving the educational value of the content.

2. Discussion

Technical and Clinical Implications. ReXVQA advances our understanding of how multimodal large language models (LLMs) reason about medical images by introducing clinically aligned evaluation dimensions such as presence detection, spatial localization, and differential diagnosis. Technically, our results reveal distinct performance patterns between generalist and specialized models. While generalist models show variable task-specific performance (excelling at negation but underperforming on geometric reasoning), the medical-specialized MedGemma demonstrates consistently high performance across all task types, highlighting the value of domain-specific training alongside granular, task-type benchmarking. Clinically, our reader studies reveal a significant milestone: MedGemma exceeds senior resident-level performance (83.24% accuracy vs. 77.27% for the best human reader), representing the first instance where AI consistently surpasses expert human evaluation in chest X-ray interpretation. The wide performance range across models from MedGemma's 83.24% to LLaVA's 24.75%, underscores the importance of careful model selection and evaluation for clinical applications. Our methodology, grounded in radiologist feedback and validated question generation, offers a scalable framework for creating trustworthy benchmarks in other medical domains, laying the groundwork for evaluating reasoning over classification in future healthcare AI systems.

2.1. *Interrater Agreement Analysis*

As shown in Figure A3, there is a clear pattern of strong human-human inter-agreement, while human-model agreement scores are comparatively lower but similar across different models.

These findings suggest that human readers share consistent interpretative frameworks and diagnostic approaches. AI models showed more variable agreement patterns, with some models like MedGemma, Qwen25VL, and Eagle2 demonstrating higher correlation with human readers and among themselves. This suggests these models may be employing reasoning patterns that more closely align with human diagnostic approaches. The All Tasks correlation matrix demonstrates that while AI models can achieve competitive individual performance, the consistency of their diagnostic reasoning across different case types remains an area for improvement. The moderate correlation coefficients between AI models and human readers (typically 0.3-0.5) indicate that despite achieving similar accuracy scores, AI models may be utilizing different diagnostic pathways than human readers.

Limitation. While ReXVQA demonstrates strong performance as a benchmark, several limitations should be acknowledged. Although we addressed demographic diversity by integrating data from four different U.S. hospital systems, the dataset may not fully represent global radiological practices or diverse international patient populations. Furthermore, our evaluation focused primarily on eight models, with limited commercial representation (only Gemini 1.5 Pro included), leaving assessment of other commercial multimodal systems as an important area for future work. While ReXVQA does employ predetermined answer choices, it differs from existing classification systems by evaluating diverse cognitive reasoning patterns rather than simple disease label prediction. Future work should explore incorporating open-ended question formats to further assess flexible clinical reasoning. These limitations present opportunities for expanding benchmark coverage across more diverse demographic settings and model architectures.

Acknowledgement

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

References

1. Alibaba DAMO Academy. Qwen-vl: An open-source vision-language model by alibaba. 2024.
2. Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
3. Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.
4. Google. Gemini pro vision: Google's multimodal model for complex reasoning. 2024.
5. Google. Medgemma hugging face. `https://huggingface.co/collections/google/medgemma-release-680aade845f90bec6a3f60c4`, 2025. Accessed: [Insert Date Accessed, e.g., 2025-05-20].
6. Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, et al. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*, 2025.
7. OBI_Lab. Clinicalbert fine-tuned for medical note de-identification. `https://huggingface.co/obi/deid_bert_i2b2`, 2022. Model fine-tuned on the I2B2 2014 dataset for PHI detection.
8. Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.
9. Microsoft Research. Phi-2 vision: Microsoft's vision-language model. 2024.
10. LLaVA Team. Llava 1.5: Open-source vision-language model. 2024.
11. Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*, 2024.

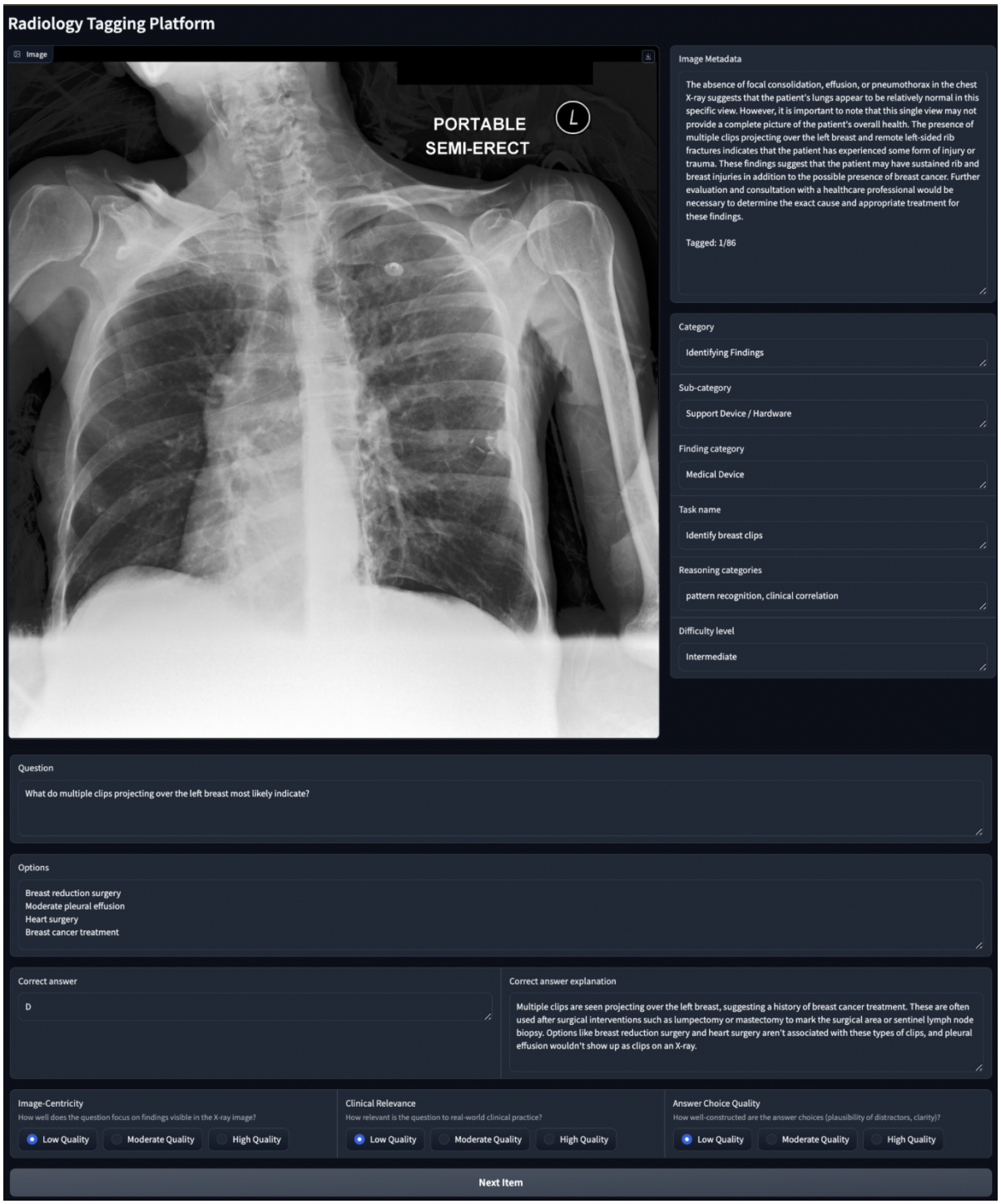

Fig. A1. **The radiology image tagging platform interface used for expert annotation.** The platform displays a portable chest X-ray with associated metadata, categorization fields, and multiple-choice assessment options. The interface includes structured fields for capturing finding categories, difficulty levels, and clinical reasoning, alongside expert feedback on image-centricity and clinical relevance. This platform facilitated systematic collection of radiologist annotations and assessments for training data validation.

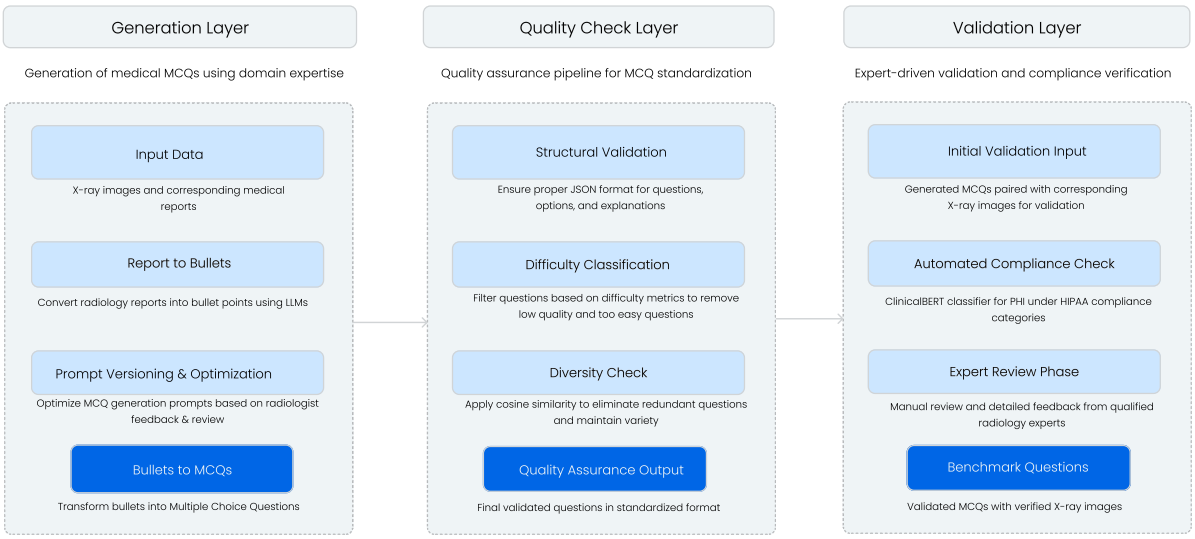

Fig. A2. **Expert-Guided Medical MCQ Generation Pipeline:** We propose a three-layer approach combining computational processes and expert oversight for creating high-quality radiology MCQs.

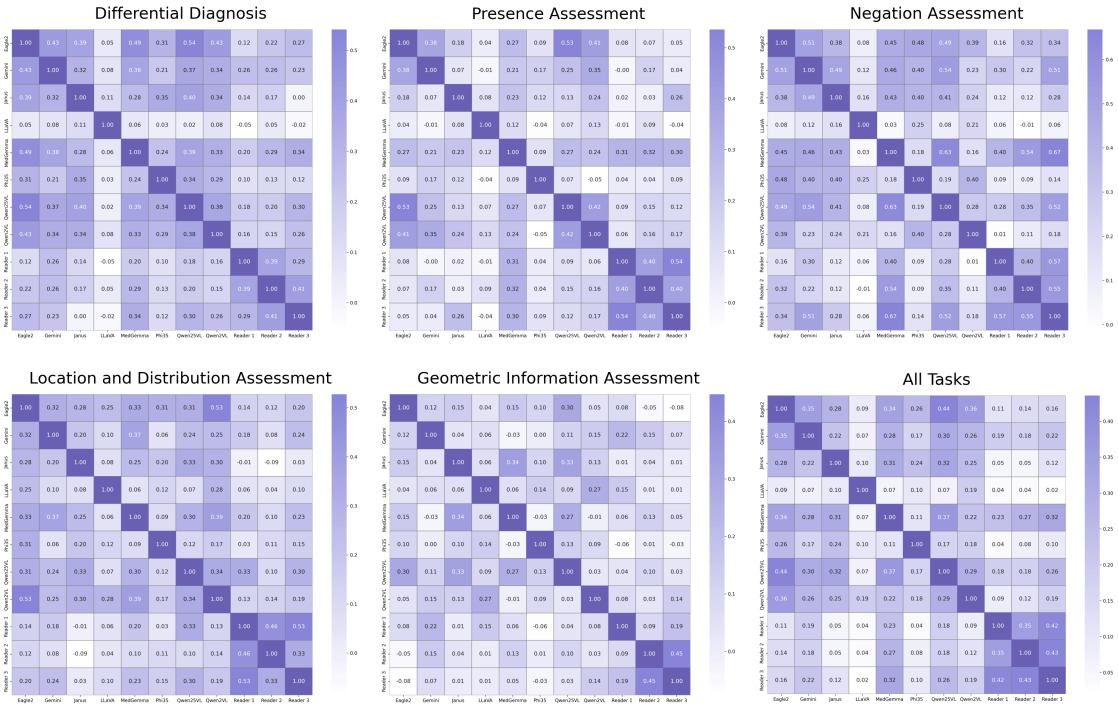

Fig. A3. Interrater Agreement Analysis Across Different Medical Assessment Tasks. We plot a heatmap showing Cohen's Kappa coefficients for interrater agreement between eight AI models and three human readers across five medical imaging assessment tasks and a combined analysis of "All Tasks". Each cell represents the kappa value between the corresponding row and column raters, with diagonal values set to 1.0 (perfect self-agreement). Purple intensity corresponds to higher agreement levels. Note that in this analysis, we include all cases from the 3 sets.