A MULTI-AGENT SYSTEM FOR COMPLEX REASONING IN RADIOLOGY VISUAL QUESTION ANSWERING

A PREPRINT

Ziruo Yi
University of North Texas
ziruoyi@my.unt.edu

Jinyu Liu
University of North Texas
jinyuliu@my.unt.edu

Ting Xiao
University of North Texas
ting.xiao@unt.edu

Mark V. Albert
University of North Texas
mark.albert@unt.edu

August 6, 2025

ABSTRACT

Radiology visual question answering (RVQA) provides precise answers to questions about chest X-ray images, alleviating radiologists' workload. While recent methods based on multimodal large language models (MLLMs) and retrieval-augmented generation (RAG) have shown promising progress in RVQA, they still face challenges in factual accuracy, hallucinations, and cross-modal misalignment. We introduce a multi-agent system (MAS) designed to support complex reasoning in RVQA, with specialized agents for context understanding, multimodal reasoning, and answer validation. We evaluate our system on a challenging RVQA set curated via model disagreement filtering, comprising consistently hard cases across multiple MLLMs. Extensive experiments demonstrate the superiority and effectiveness of our system over strong MLLM baselines, with a case study illustrating its reliability and interpretability. This work highlights the potential of multi-agent approaches to support explainable and trustworthy clinical AI applications that require complex reasoning.

Keywords Radiology Visual Question Answering · Complex Reasoning · Multimodal Large Language Models · Multi-Agent Systems · Retrieval-Augmented Generation

1 Introduction

In modern healthcare, radiology plays a crucial role in diagnosis, outcome prediction, and treatment planning. It relies on diverse data sources such as chest X-ray images, clinical notes, and laboratory tests. Among various multimodal tasks in radiology, radiology visual question answering (RVQA) is particularly valuable as it provides accurate answers to questions about chest X-ray images [1] and reduces repetitive tasks, thereby alleviating radiologists' workload [2]. However, many existing RVQA benchmarks focus on relatively simple question formats and reasoning types. For example, VQA-RAD [3] and Slake [4] primarily consist of Yes/No or open-ended questions, with limited diversity in task types or reasoning complexity. Although newer datasets such as EHRXQA [5] offer larger scale and more diverse question templates, the lack of expert-level answer explanations limits their ability to evaluate models' complex reasoning capabilities. While existing benchmarks mainly contain questions requiring only basic reasoning, real-world RVQA scenarios involve subtle visual cues, multi-step inference, and domain knowledge integration, which present significant challenges to the complex reasoning capabilities of current models. This gap underscores the urgent need for RVQA systems that can handle complex, clinically aligned reasoning over multimodal inputs.

With recent advancements in artificial intelligence (AI), computer vision (CV), and natural language processing (NLP), multimodal learning has emerged as a powerful paradigm for integrating and analyzing diverse data sources [6, 7]. Powered by large language models (LLMs) and large vision models (LVMs) such as GPT-4o [8], LLaMA 3 [9], and DALL·E 3 [10], recent multimodal large language models (MLLMs) have shown promising results across a range of tasks including image captioning [11] and visual-language dialogue [12]. In the healthcare domain, MLLMs such as Med-PaLM 2 [13] and LLaVA-Med [14] have made notable progress in pharmaceutical research [15] and clinical support [16]. In RVQA, MLLMs integrate visual and textual information to provide precise answers that support clinical decision-making [17, 18]. However, existing MLLM approaches for RVQA often treat the model as a black

box, combining visual understanding, language grounding, and answer synthesis into a single step. This unified design makes it difficult to interpret intermediate steps, increases the risk of hallucinations under ambiguous cases, and limits the model's ability to handle complex diagnostic scenarios.

Retrieval-augmented generation (RAG) [19, 20, 21] has emerged as a promising strategy to improve the factual accuracy of medical MLLMs. By incorporating external knowledge, RAG enhances contextual understanding and enables more grounded responses. It has been applied to various medical tasks, including report generation [22, 23] and visual question answering (VQA) [24]. However, applying RAG to RVQA presents new challenges. Retrieving too many contexts can introduce noise and redundancy, while insufficient retrieval may miss key information, ultimately reducing overall answer quality. Given the complexity of RVQA, which demands reliable retrieval, medical reasoning, and cross-modal integration, a single model often struggles to fulfill all requirements. Recently, LLM-based agentic systems have attracted increasing interest due to their ability to perform complex, multi-step tasks through structured collaboration and dynamic interaction [25, 26]. Building on this foundation, multi-agent systems (MASs) have shown potential in areas such as software engineering [27, 28] and drug discovery [29, 30]. However, their application in radiology, especially in RVQA, remains underexplored.

To overcome the limitations of existing MLLM and RAG approaches and to explore the potential of multi-agent systems in radiology, we propose a multi-agent system (MAS) composed of three specialized agents for RVQA: a context understanding agent (CUA), a multimodal reasoning agent (MRA), and an answer validation agent (AVA). This modular design enables structured, stepwise collaboration among agents, enhancing the explainability and precision of the reasoning process while reducing hallucinations. To support our MAS and comprehensively evaluate its complex reasoning capabilities, we curate three subsets from ReXVQA [31], a large-scale benchmark of multiple-choice questions covering diverse radiological tasks and categories. These subsets enable model disagreement analysis, serve as a retrieval pool for RAG, and facilitate evaluation on challenging cases. The main contributions of our work are as follows: (1) We present a modular MAS for RVQA that leverages role-specific agents to collaboratively perform complex reasoning tasks, including answer prediction and explanation generation. (2) We construct three subsets based on ReXVQA to support our MAS and systematically evaluate its performance in challenging scenarios. (3) We conduct extensive experiments and analyses showing that our MAS consistently outperforms strong MLLM baselines in accuracy, interpretability, and robustness on ambiguous cases.

2 Related Work

MLLMs for RVQA. MLLMs have made significant progress in RVQA [32, 33, 34, 35, 36, 37], with models such as ELIXR [38], LLaVa-Med [14], and PeFoMed [39] combining visual encoders (e.g., ViT [40], EVA [41]) with LLMs (e.g., Vicuna [42], LLaMA2-chat [43]) to align visual features with textual representations and achieve strong performance on benchmark datasets. Despite their success, these models still suffer from key limitations including reasoning inconsistency [44, 45], hallucination [46], and catastrophic forgetting [47, 48]. These issues are especially critical in RVQA, where high accuracy and reliability are essential [49]. In contrast to these black-box approaches, our MAS adopts a modular decomposition with explicit retrieval and validation stages, enhancing reasoning control, interpretability, and robustness.

Retrieval-Augmented Generation. RAG has been increasingly adopted to improve factual grounding in multimodal tasks by retrieving relevant external knowledge [50, 51]. It has been applied to RVQA to reduce hallucinations and enhance factual accuracy [52]. While RAG offers clear benefits, it still struggles to manage the quantity and quality of retrieved contexts and to mitigate over-reliance on them, which may compromise model performance and introduce factual errors [53]. Moreover, many RAG methods retrieve and process textual and visual information separately, limiting their capacity for integrated multimodal reasoning [54]. These limitations are particularly critical in RVQA, where accurate reasoning requires fine-grained alignment between retrieved information and visual evidence. To address these challenges, our system incorporates RAG into a multi-agent workflow, where a dedicated agent retrieves semantically relevant references to support downstream reasoning and validation using both textual and visual inputs. This design enables stronger factual grounding and cross-modal alignment, making it better suited to the complex reasoning demands of RVQA.

Multi-Agent Systems. MASs have gained increasing attention in NLP and healthcare AI [55, 56, 57, 58, 59]. Typically, a MAS consists of a collection of agents that interact through orchestration to enable collective intelligence via coordinated task decomposition, performance parallelization, context isolation, specialized model ensembling, and diverse reasoning discussions [60, 61, 62, 63, 64, 65, 66]. MASs distribute tasks among specialized agents to collaboratively accomplish complex objectives beyond the scope of single models. While MASs have shown promising results in radiology report generation (RRG) [67, 68], their application to RVQA remains largely unexplored. To fill this gap, we propose a multimodal MAS for RVQA that decomposes the complex reasoning process into structured

subtasks handled by three specialized agents: a CUA, a MRA, and an AVA. This design facilitates transparent, step-wise collaboration and enables more accurate, robust, and clinically aligned reasoning required in complex radiological scenarios.

3 Method

We propose a modular multi-agent system for RVQA that decomposes complex multimodal reasoning into interpretable and cooperative stages. Given a multiple-choice question (MCQ) and one or more corresponding X-ray images, the system sequentially activates three specialized agents: a CUA, a MRA, and an AVA. Each agent serves a distinct functional role and operates independently, using either task-specific prompts (for LLM/MLLM agents) or embedding-based retrieval for selecting top-k relevant examples. As illustrated in Figure 1, the agents communicate through structured intermediate outputs that progressively refine both context understanding and multimodal reasoning into a final answer and explanation. This stepwise architecture enables modular design, interpretability, and flexible integration of different LLMs and MLLMs across agents, aligning with the needs of complex reasoning in radiology MCQs.

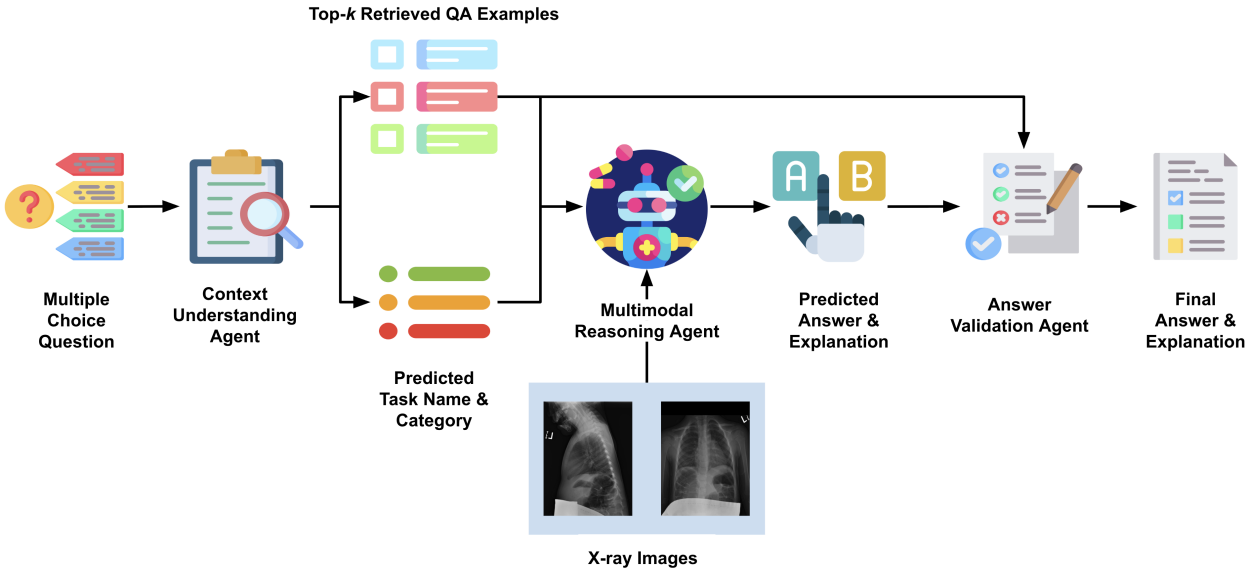

Figure 1: Overview of the proposed multi-agent system. The pipeline consists of: (1) a context understanding agent that retrieves top-k relevant QA examples and predicts task name and category; (2) a multimodal reasoning agent that generates image-grounded answers and explanations; and (3) an answer validation agent that verifies the predicted answer and revises it when necessary.

3.1 Context Understanding Agent

The CUA supports downstream reasoning by retrieving top-k relevant QA examples and identifying the radiological reasoning task and category of each question. Given an MCQ, we first apply an embedding-based retriever to obtain the top-n most similar QA examples based on sentence-level similarity. These candidates are then reranked by an LLM that scores their relevance to the input MCQ, considering semantic similarity, clinical relevance, and reasoning alignment. We predict the task name and diagnostic category of the input MCQ via weighted voting over the top-k reranked examples. This step provides the predicted task name, category, and top-k QA examples most relevant to the input, which together serve as contextual inputs for the next agent.

3.2 Multimodal Reasoning Agent

The MRA generates an answer and explanation by integrating the MCQ, its corresponding X-ray image(s), top-k QA examples, and the predicted task name and category. These elements are passed into a MLLM that leverages both visual evidence and textual knowledge. The MRA then produces a selected answer (A/B/C/D) and a free-text

explanation for the choice. This agent produces both the predicted answer and an explanation, together with the top-k QA examples used as inputs for the answer validation step.

3.3 Answer Validation Agent

The AVA assesses the reliability of the predicted answer and performs correction if its confidence score falls below a predefined threshold. It receives the MCQ, the top-k QA examples retrieved by the CUA, and the predicted answer and explanation generated by the MRA. The AVA uses an LLM to estimate the confidence score of the predicted answer. If the score exceeds the threshold, the predicted answer and explanation are accepted. Otherwise, the LLM reconsiders the MCQ using the top-k QA examples and generates a revised answer and explanation. This process produces the final output, improving overall answer reliability by correcting low-confidence predictions when necessary.

4 Experiments

We conduct a comprehensive evaluation of our multi-agent system on ReXVQA-Hard, a challenging test set from the ReXVQA dataset. Our experiments assess the pipeline's performance in answer accuracy and explanation quality, compare it with strong MLLM baselines, and analyze the contributions of individual agents to complex reasoning in RVQA.

4.1 Experimental Setup

4.1.1 Implementation Details.

Our system consists of three agents: the CUA, MRA and AVA. The CUA employs Facebook AI Similarity Search (FAISS) [69] to retrieve the top-n relevant QA examples ($n = 10$), which are subsequently reranked by MMed-Llama-3-8B [70]. We apply a rule-based weighted voting over the top-k examples ($k = 5$) to predict the task name and radiological category, assigning greater weight to higher-ranked examples. The MRA is implemented using Med-GEMMA [71], an MLLM fine-tuned for medical image and language understanding. MMed-Llama-3-8B is also employed in the AVA for answer validation and correction, using a confidence threshold of 0.7 selected for its consistently strong performance across different values. Since GPT-4o was used to generate the reference answers in the dataset, we exclude GPT-based models from all components of our system to prevent evaluation bias and ensure fairness.

4.1.2 Datasets.

We conduct experiments on ReXVQA, a large-scale benchmark for MCQs in chest radiology. Each example includes one or more chest X-ray images, a clinically meaningful question with four answer options (A/B/C/D), and an expert-written explanation. The dataset covers five radiological reasoning tasks (e.g., presence assessment, differential diagnosis), spanning a diverse range of clinical categories.

ReXVQA-Pool. We randomly select 600 examples per task, forming a 3,000-example subset that reflects diverse radiological categories. This subset contains a total of 4,795 chest X-ray images and is used to support model disagreement analysis across existing MLLMs.

ReXVQA-Hard. We identify 1,131 challenging examples from ReXVQA-Pool on which most MLLMs fail to answer correctly, forming a hard subset that serves as the test set for evaluating our multi-agent system.

ReXVQA-RAG. We construct a separate QA bank of 1,006 diverse examples from the portion of ReXVQA not included in ReXVQA-Pool. This subset is used by the CUA during the retrieval step.

4.1.3 Baseline Models.

To evaluate the effectiveness of our multi-agent system, we compare its performance against several state-of-the-art (SOTA) MLLMs on ReXVQA-Hard. These include general-purpose MLLMs (e.g., Janus-Pro-7B [72], LLaVA 1.5 [73], OpenFlamingo-4B [74], Phi-3.5-Vision-Instruct [75], and Qwen2.5-VL [76]) as well as a medical-domain MLLM, MedGemma. The selected baselines span a diverse range of architectures and training strategies, providing a representative view of current MLLM capabilities. Consistent with our system, we exclude all GPT-based models from the baselines to avoid evaluation bias, since GPT-4o was used to generate the dataset's reference answers. To improve clarity, we refer to the baseline models using the following short names throughout the paper: MedGemma, Janus (Janus-Pro-7B), LLaVA (LLaVA 1.5), OpenFlamingo (OpenFlamingo-4B), Phi35 (Phi-3.5-Vision-Instruct), and

Qwen25VL (Qwen2.5-VL). Although Phi-3.5-Vision did not participate in the construction of the ReXVQA-Hard subset, we include it in our evaluation to enable broader comparison across diverse MLLMs.

4.1.4 Evaluation Metrics.

The performance of our MAS is evaluated using accuracy for answer prediction and standard text generation metrics including BLEU [77], ROUGE-L [78], METEOR [79], and BERTScore [80] for explanation quality. These metrics are designed to capture both lexical and semantic similarity between the generated explanations and the ground-truth references, ranging from surface-level overlap (e.g., BLEU, ROUGE-L) to deeper alignment (e.g., METEOR, BERTScore).

4.1.5 Construction of ReXVQA-Hard

To construct the challenging test set ReXVQA-Hard, we adopt a difficulty-based filtering strategy inspired by MedAgentsBench [81], which highlights the importance of evaluating the complex medical reasoning capabilities in current models. Specifically, we run five MLLMs including MedGemma, Janus, LLaVA, OpenFlamingo, and Qwen25VL on ReXVQA-Pool, and select examples that are incorrectly answered by at least three of them. This approach ensures that ReXVQA-Hard contains consistently difficult examples across diverse model architectures. As shown in Table 1, the five MLLMs exhibit substantial variation in accuracy on ReXVQA-Pool, ranging from 31.87% (LLaVA) to 70.77% (MedGemma). This performance gap suggests that many examples are answered correctly by some models and incorrectly by others, enabling meaningful disagreement-based filtering. By selecting examples that are incorrectly answered by at least three models, we capture questions that are not only difficult for one specific architecture but consistently challenging across models with diverse training approaches. This variation also highlights the limitations of general-purpose MLLMs in handling clinically grounded reasoning tasks, further motivating the need for focused evaluation on hard cases.

Table 1: Accuracy of different MLLMs on ReXVQA-Pool used to construct ReXVQA-Hard.

Metric	**MedGemma**	**Janus**	**LLaVA**	**OpenFlamingo**	**Qwen25VL**
Accuracy (%)	70.77	65.30	31.87	42.40	68.27

4.2 Results

To evaluate the effectiveness of our MAS, we conduct both quantitative and qualitative analyses on ReXVQA-Hard. This section compares our system with several strong baseline MLLMs in terms of answer accuracy and explanation quality, and presents a case study illustrating how the agents collaborate to resolve diagnostic ambiguity. These results provide a comprehensive assessment of the pipeline's ability to handle complex reasoning in RVQA.

4.2.1 Quantitative Analysis

We evaluate our MAS on ReXVQA-Hard using standard metrics for both answer prediction and explanation generation. As shown in Table 2, our approach consistently and significantly outperforms recent SOTA MLLMs, including MedGemma, Janus, LLaVA, OpenFlamingo, Phi35, and Qwen25VL. Specifically, our method achieves an accuracy of 63.66%, outperforming the strongest baseline, MedGemma (44.03%), by nearly 20 percentage points. This notable improvement highlights our system's enhanced ability to handle complex and ambiguous examples. In terms of explanation quality, our system obtains the highest scores in BLEU (0.1230) and ROUGE-L (0.3692), suggesting improved lexical alignment with reference explanations. For METEOR, our model achieves a strong score of 0.3449, which is competitive with the best-performing Qwen25VL (0.4125). Our pipeline also achieves a BERTScore of 0.8987, which is comparable to the top score of 0.9008 from Phi35 and reflects strong semantic consistency. These results highlight the effectiveness of integrating contextual understanding, visual reasoning, and answer validation within a MAS. By jointly optimizing for both accuracy and explanation quality, our system generates responses that are not only correct but also clinically meaningful. This makes it particularly suitable for solving difficult MCQs that require complex reasoning in RVQA.

4.2.2 Qualitative Analysis

Figure 2 illustrates how our MAS addresses a challenging MCQ. It coordinates contextual understanding, multimodal reasoning, and answer validation to arrive at the correct prediction. The input MCQ asks for the most likely con-

Table 2: Quantitative comparison between our MAS and SOTA MLLMs on ReXVQA-Hard.

Model	Accuracy	BLEU	ROUGE-L	METEOR	BERTScore
MedGemma	44.03%	0.0421	0.2071	0.1903	0.8755
Janus	35.46%	0.0192	0.2118	0.2536	0.8691
LLaVA	26.08%	0.0048	0.0407	0.0521	0.8171
OpenFlamingo	28.12%	0.0612	0.2567	0.2741	0.8798
Phi35	27.94%	0.0802	0.3325	0.3130	**0.9008**
Qwen25VL	29.53%	0.0750	0.2977	**0.4125**	0.8912
Ours	**63.66%**	**0.1230**	**0.3692**	0.3449	0.8987

dition based on a chest X-ray, with options including congestive heart failure and pulmonary embolism. The CUA correctly identifies the radiological task name and category, and retrieves QA examples with similar questions and closely related options such as congestive heart failure. The MRA integrates visual and contextual cues to generate a plausible explanation, but ultimately selects an incorrect answer ("C. Pulmonary embolism"), possibly due to the subtle appearance of heart failure on the image. The AVA estimates the confidence of this prediction using an LLM, and determines that it falls below the predefined threshold. It then reconsiders the question using the retrieved QA examples and produces a revised answer: "B. Congestive heart failure," accompanied by a clinically aligned explanation that supports the diagnosis. This case highlights how contextual knowledge, visual information, and answer validation jointly contribute to resolving diagnostic ambiguity and improving reasoning accuracy.

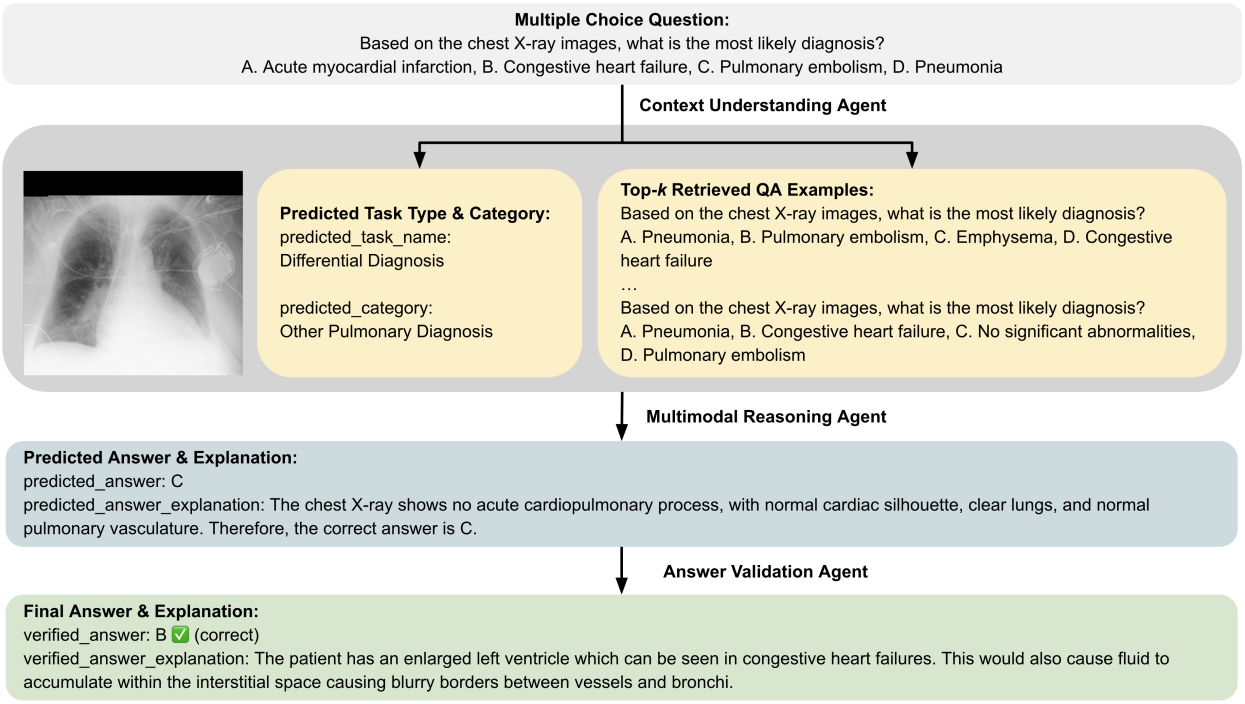

Figure 2: A case study showing how the multi-agent system integrates contextual information, visual evidence, and answer validation to support accurate diagnostic reasoning.

4.3 Ablation Study

Agent Contributions to Overall Performance

To analyze the contribution of each agent, we conduct an ablation study by incrementally enabling them in the pipeline. Specifically, we compare the following three configurations:

1. **MRA only**: The MRA generates answers and explanations using only the input MCQ and its associated X-ray image(s), without access to retrieved examples or any contextual information.

2. **CUA + MRA**: The CUA provides retrieved QA examples along with the predicted task name and category, which are used by the MRA in addition to the input MCQ and image(s) to generate an answer and explanation. The AVA is disabled in this configuration, so no verification or correction is applied.

3. **Full system (CUA + MRA + AVA)**: In the full configuration, the AVA estimates the confidence score of the predicted answer. If the score falls below a predefined threshold, it reconsiders the question using the retrieved QA examples and generates a revised answer and explanation.

The performance of each configuration is shown in Table 3, with accuracy evaluating answer correctness and BLEU, ROUGE-L, METEOR, and BERTScore assessing explanation quality. Using only the MRA results in the lowest performance across all metrics (e.g., accuracy: 44.03%, BLEU: 0.0421, BERTScore: 0.8755), indicating that image and question inputs alone are insufficient for reliable reasoning or explanation generation. The lack of contextual guidance limits the model's ability to produce clinically meaningful outputs. Introducing the CUA substantially boosts performance. Accuracy increases by over 10 percentage points (from 44.03% to 54.29%), and explanation metrics see similar gains (e.g., BLEU increases from 0.0421 to 0.1078, ROUGE-L from 0.2071 to 0.3532). These results demonstrate that predicted task name, category, and retrieved QA examples are highly effective in guiding both answer selection and explanation generation. Enabling the full system with the AVA further improves performance, achieving 63.66% accuracy and the highest explanation scores across all metrics (e.g., BLEU: 0.1230, BERTScore: 0.8987). These improvements highlight the value of answer validation in identifying and correcting low-confidence predictions, ultimately improving output accuracy and trustworthiness. Overall, each component plays a critical role in system performance. The full pipeline benefits from contextual retrieval, multimodal reasoning, and answer validation, highlighting the importance of a modular and cooperative architecture for complex reasoning in RVQA.

Table 3: Performance comparison of different agent configurations on ReXVQA-Hard.

Agents	Accuracy	BLEU	ROUGE-L	METEOR	BERTScore
MRA only	44.03%	0.0421	0.2071	0.1903	0.8755
CUA + MRA	54.29%	0.1078	0.3532	0.3225	0.8950
Full System	**63.66%**	**0.1230**	**0.3692**	**0.3449**	**0.8987**

4.4 Discussion

Our experiments demonstrate that the proposed MAS significantly improves performance on ReXVQA-Hard, a test set selected based on model disagreement to include consistently difficult examples. By combining contextual retrieval, multimodal reasoning, and answer validation, our system consistently outperforms strong baselines across both answer accuracy and explanation quality. These results highlight the importance of modular and interpretable reasoning pipelines for tackling radiological questions that require complex reasoning.

The CUA enriches contextual information by retrieving semantically relevant QA examples and predicting the radiological task type and category, which provides critical guidance to downstream components. The MRA then fuses visual and textual modalities to generate clinically plausible answers and explanations, effectively addressing the complexity of radiological reasoning. Finally, the AVA estimates the confidence of each prediction and revises low-confidence answers when necessary, improving reliability on clinically ambiguous cases.

The integration of the three specialized agents results in a modular system that balances performance, explainability, and clinical alignment. The ablation study confirms that each agent plays a key role in the system's overall performance, while the case study illustrates how their coordination improves diagnostic reasoning in challenging cases. One limitation is that the AVA relies on a fixed confidence threshold to trigger answer revision, which may not generalize well across different medical domains. To address this, future work will explore dynamic thresholding strategies or learned validation mechanisms to enhance flexibility and robustness, especially for difficult or rare cases.

5 Conclusion

We present a MAS for RVQA, which decomposes the task into three specialized agents for context understanding, multimodal reasoning, and answer validation. This modular design fully leverages multimodal information and enables interpretable and step-wise reasoning. Our system significantly outperforms strong MLLM baselines, achieving notable improvements in both answer accuracy and explanation quality. The ablation study demonstrates the importance of each agent, while qualitative analysis illustrates how their collaboration resolves diagnostic ambiguity.

This pipeline provides a flexible and generalizable approach for other multimodal medical tasks that require complex reasoning and clinical precision.

References

[1] Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in artificial intelligence*, 7:1430984, 2024.

[2] Ziruo Yi, Ting Xiao, and Mark V Albert. A survey on multimodal large language models in radiology for report generation and visual question answering. *Information*, 16(2):136, 2025.

[3] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

[4] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[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*, 36:3867–3880, 2023.

[6] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021.

[7] Asim Waqas, Aakash Tripathi, Ravi P Ramachandran, Paul A Stewart, and Ghulam Rasool. Multimodal data integration for oncology in the era of deep neural networks: a review. *Frontiers in Artificial Intelligence*, 7:1408843, 2024.

[8] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[9] AI Meta. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*, 2024.

[10] OpenAI. DALL-E3, 2023. `https://openai.com/index/dall-e-3/`.

[11] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[12] Yupan Huang, Zaiqiao Meng, Fangyu Liu, Yixuan Su, Nigel Collier, and Yutong Lu. Sparkles: Unlocking chats across multiple images for multimodal instruction-following models. *arXiv preprint arXiv:2308.16463*, 2023.

[13] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8, 2025.

[14] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

[15] Francesca Grisoni. Chemical language models for de novo drug design: Challenges and opportunities. *Current Opinion in Structural Biology*, 79:102527, 2023.

[16] Stephen R Ali, Thomas D Dobbs, Hayley A Hutchings, and Iain S Whitaker. Using chatgpt to write patient clinic letters. *The Lancet Digital Health*, 5(4):e179–e181, 2023.

[17] Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*, 2023.

[18] Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Laila Bashmal, and Mansour Zuair. Vision–language model for visual question answering in medical imagery. *Bioengineering*, 10(3):380, 2023.

[19] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1, 2023.

[20] Xiaoye Qu, Qiyuan Chen, Wei Wei, Jishuo Sun, and Jianfeng Dong. Alleviating hallucination in large vision-language models with active retrieval augmentation. *arXiv preprint arXiv:2408.00555*, 2024.

[21] Xiaoye Qu, Jiashuo Sun, Wei Wei, and Yu Cheng. Look, compare, decide: Alleviating hallucination in large vision-language models via multi-view multi-path reasoning. *arXiv preprint arXiv:2408.17150*, 2024.

[22] Yogesh Kumar and Pekka Marttinen. Improving medical multi-modal contrastive learning with expert annotations. In *European Conference on Computer Vision*, pages 468–486. Springer, 2024.

[23] Yitian Tao, Liyan Ma, Jing Yu, and Han Zhang. Memory-based cross-modal semantic alignment network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[24] Zheng Yuan, Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Hongyi Yuan, Fei Huang, and Songfang Huang. Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 547–556, 2023.

[25] Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. Gorilla: Large language model connected with massive apis. *Advances in Neural Information Processing Systems*, 37:126544–126565, 2024.

[26] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

[27] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. Chatdev: Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 2023.

[28] Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. Openhands: An open platform for ai software developers as generalist agents. *arXiv preprint arXiv:2407.16741*, 2024.

[29] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

[30] Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation. *bioRxiv*, pages 2024–11, 2024.

[31] Ankit Pal, Jung-Oh Lee, Xiaoman Zhang, Malaikannan Sankarasubbu, Seunghyeon Roh, Won Jung Kim, Meesun Lee, and Pranav Rajpurkar. Rexvqa: A large-scale visual question answering benchmark for generalist chest x-ray understanding. *arXiv preprint arXiv:2506.04353*, 2025.

[32] Cuong Nhat Ha, Shima Asaadi, Sanjeev Kumar Karn, Oladimeji Farri, Tobias Heimann, and Thomas Runkler. Fusion of domain-adapted vision and language models for medical visual question answering. *arXiv preprint arXiv:2404.16192*, 2024.

[33] Pengfei Li, Gang Liu, Jinlong He, Zixu Zhao, and Shenjun Zhong. Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 374–383. Springer, 2023.

[34] Tiancheng Gu, Kaicheng Yang, Dongnan Liu, and Weidong Cai. Lapa: Latent prompt assist model for medical visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2024.

[35] Timothy Ossowski and Junjie Hu. Multimodal prompt retrieval for generative visual question answering. *arXiv preprint arXiv:2306.17675*, 2023.

[36] Jonggwon Park, Soobum Kim, Byungmu Yoon, Jihun Hyun, and Kyoyun Choi. M4cxr: Exploring multi-task potentials of multi-modal large language models for chest x-ray interpretation. *arXiv preprint arXiv:2408.16213*, 2024.

[37] Taehee Kim, Yeongjae Cho, Heejun Shin, Yohan Jo, and Dongmyung Shin. Generalizing visual question answering from synthetic to human-written questions via a chain of qa with a large language model. In *ECAI 2024*, pages 298–305. IOS Press, 2024.

[38] Shawn Xu, Lin Yang, Christopher Kelly, Marcin Sieniek, Timo Kohlberger, Martin Ma, Wei-Hung Weng, Atilla Kiraly, Sahar Kazemzadeh, Zakkai Melamed, et al. Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders. *arXiv preprint arXiv:2308.01317*, 2023.

[39] Gang Liu, Jinlong He, Pengfei Li, Genrong He, Zhaolin Chen, and Shenjun Zhong. Pefomed: Parameter efficient fine-tuning of multimodal large language models for medical imaging. *arXiv preprint arXiv:2401.02797*, 2024.

[40] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[41] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19358–19369, 2023.

[42] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, march 2023. *URL https://lmsys. org/blog/2023-03-30-vicuna*, 3(5), 2023.

[43] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[44] Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. *arXiv preprint arXiv:2408.12076*, 2024.

[45] Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *Advances in Neural Information Processing Systems*, 37:140334–140365, 2024.

[46] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024.

[47] Hikmat Khan, Nidhal C Bouaynaya, and Ghulam Rasool. The importance of robust features in mitigating catastrophic forgetting. In *2023 IEEE Symposium on Computers and Communications (ISCC)*, pages 752–757. IEEE, 2023.

[48] Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023.

[49] Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*, 2025.

[50] Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *arXiv preprint arXiv:2502.08826*, 2025.

[51] Matin Mortaheb, Mohammad A Amir Khojastepour, Srimat T Chakradhar, and Sennur Ulukus. Re-ranking the context for multimodal retrieval augmented generation. *arXiv preprint arXiv:2501.04695*, 2025.

[52] Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*, 2024.

[53] Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. Rule: Reliable multimodal rag for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1093, 2024.

[54] Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. Mdocagent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964*, 2025.

[55] Yu He Ke, Rui Yang, Sui An Lie, Taylor Xin Yi Lim, Hairil Rizal Abdullah, Daniel Shu Wei Ting, and Nan Liu. Enhancing diagnostic accuracy through multi-agent conversations: using large language models to mitigate cognitive bias. *arXiv preprint arXiv:2401.14589*, 2024.

[56] Andries Petrus Smit, Paul Duckworth, Nathan Grinsztajn, Kale-ab Tessera, Thomas D Barrett, and Arnu Pretorius. Are we going mad? benchmarking multi-agent debate between language models for medical q&a. In *Deep Generative Models for Health Workshop NeurIPS 2023*, 2023.

[57] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*, 2023.

[58] Hao Wei, Jianing Qiu, Haibao Yu, and Wu Yuan. Medco: Medical education copilots based on a multi-agent framework. *arXiv preprint arXiv:2408.12496*, 2024.

[59] Ling Yue and Tianfan Fu. Ct-agent: Clinical trial multi-agent with large language model-based reasoning. *arXiv e-prints*, pages arXiv–2404, 2024.

[60] Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, et al. Why do multi-agent llm systems fail? *arXiv preprint arXiv:2503.13657*, 2025.

[61] Junda He, Christoph Treude, and David Lo. Llm-based multi-agent systems for software engineering: Literature review, vision, and the road ahead. *ACM Transactions on Software Engineering and Methodology*, 34(5):1–30, 2025.

[62] Zhao Mandi, Shreeya Jain, and Shuran Song. Roco: Dialectic multi-robot collaboration with large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 286–299. IEEE, 2024.

[63] Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485*, 2023.

[64] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

[65] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

[66] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

[67] Hasan Md Tusfiqur Alam, Devansh Srivastav, Md Abdul Kadir, and Daniel Sonntag. Towards interpretable radiology report generation via concept bottlenecks using a multi-agentic rag. In *European Conference on Information Retrieval*, pages 201–209. Springer, 2025.

[68] Fang Zeng, Zhiliang Lyu, Quanzheng Li, and Xiang Li. Enhancing llms for impression generation in radiology reports through a multi-agent system. *arXiv preprint arXiv:2412.06828*, 2024.

[69] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[70] Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384, 2024.

[71] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[72] 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.

[73] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[74] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[75] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[76] 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.

[77] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[78] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[79] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[80] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[81] Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, et al. Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning. *arXiv preprint arXiv:2503.07459*, 2025.