AMANDA: Agentic Medical Knowledge Augmentation for Data-Efficient Medical Visual Question Answering

Anonymous ACL submission

Abstract

Medical Multimodal Large Language Models (Med-MLLMs) have shown great promise in medical visual question answering (Med-VQA). However, when deployed in low-resource settings where abundant labeled data are unavailable, existing Med-MLLMs commonly fail due to their medical reasoning capability bottlenecks: (i) the intrinsic reasoning bottleneck that ignores the details from the medical image; (ii) the extrinsic reasoning bottleneck that fails to incorporate specialized medical knowledge. To address those limitations, we propose **AMANDA**, a training-free agentic framework that performs medical knowledge augmentation via LLM agents. Specifically, our intrinsic medical knowledge augmentation focuses on coarse-to-fine question decomposition for comprehensive diagnosis, while extrinsic medical knowledge augmentation grounds the reasoning process via biomedical knowledge graph retrieval. Extensive experiments across eight Med-VQA benchmarks demonstrate substantial improvements in both zero-shot and few-shot Med-VQA settings. The code is available at `https://anonymous.4open.science/r/AMANDA-CF56`.

1 Introduction

Medical Visual Question Answering (Med-VQA) aims to automatically answer natural language questions about medical images, serving as an AI-powered assistant to enhance healthcare professionals' diagnostic efficiency and accuracy (Hartsock and Rasool, 2024; Lin et al., 2023b). Unlike general-domain VQA which focuses on everyday scenes and objects, Med-VQA requires fine-grained analysis of subtle pathological features, understanding of professional medical terminology, and integration of domain-specific medical knowledge (Lin et al., 2023b). These unique characteristics make Med-VQA particularly challenging yet crucial for empowering precise medical diagnosis.

Recent advances in Medical Multimodal Large Language Models (Med-MLLMs) have demonstrated promising results in Med-VQA through extensive pre-training and task-specific fine-tuning (Li et al., 2024b; Eslami et al., 2023; Zhang et al., 2023b; Jiang et al., 2024c). However, obtaining a large-scale medical dataset for Med-MLLM pre-training or fine-tuning requires labor-intensive expert annotations, making it impractical in data-efficient scenarios. When deployed in low-resource settings where abundant training or fine-tuning data are unavailable (i.e., zero-shot or few-shot settings), existing Med-MLLMs commonly fail due to two bottlenecks of their medical reasoning capability:

- From the *intrinsic* perspective, current Med-MLLMs usually focus on understanding the image from a general view, while ignoring the fine-grained examination of subtle pathological features that are critical for accurate diagnosis (Lin et al., 2023b). In clinical practice, medical professionals achieve comprehensive analysis through an iterative process of questioning and examination, progressively uncovering crucial details. However, the single-step inference adopted by existing Med-MLLMs fails to capture this iterative nature of the medical diagnosis, leading to superficial analyses without critical diagnostic details (Wang et al., 2023; Jiang et al., 2024a,b).
- From the *extrinsic* perspective, while Med-MLLMs possess basic medical knowledge through pre-training, these models are typically static and lack mechanisms to access or incorporate new medical knowledge continually. In Med-VQA tasks, such specialized medical knowledge from up-to-date knowledge bases is particularly crucial. Correspondingly, existing methods often struggle to provide comprehensive and contextually grounded answers, with a concerning tendency to generate hallucinations (Xia et al., 2024b; Yan et al., 2024) – plausible but factually

incorrect responses that pose significant risks for real-world medical diagnosis.

To address the aforementioned challenges, we present a training-free MLLM agentic framework – AMANDA (Agentic MedicAl KNowleDge Augmentation) for data-efficient medical visual question answering. In essence, our framework enhances Med-MLLMs' reasoning capability through *Medical Knowledge Augmentation* (Med-KA) from both intrinsic and extrinsic reasoning perspectives. On the one hand, to enhance the medical reasoning depth, we propose *Intrinsic Med-KA*, which leverages a coarse-to-fine question decomposition strategy to fully utilize the intrinsic visual understanding capabilities within Med-MLLMs, enabling comprehensive diagnosis through progressive examination. On the other hand, to bridge the gap between models' pre-trained knowledge and reliable medical expertise, we develop *Extrinsic Med-KA*, which retrieves relevant medical knowledge from biomedical knowledge graphs to ground the reasoning process. These complementary approaches are orchestrated by multiple LLM agents that can adaptively control the depth of knowledge integration to maintain both effectiveness and efficiency. In addition, AMANDA can incorporate in-context learning examples, enabling further performance gains in few-shot settings. Overall, our contributions can be summarized as follows:

- **Problem.** We target the challenging problem of data-efficient Med-VQA and propose a training-free agentic framework that addresses the intrinsic and extrinsic bottlenecks of Med-MLLMs' reasoning capability via Med-KA.
- **Method.** We develop a Med-KA approach from two complementary perspectives: *intrinsic Med-KA* through coarse-to-fine question decomposition and *extrinsic Med-KA* via medical knowledge graph retrieval, unified under an adaptive refinement mechanism.
- **Experiments.** Through comprehensive experiments on eight Med-VQA benchmarks, we demonstrate substantial improvements in both zero-shot and few-shot settings, with strong generalization across different types of MLLMs.

2 Related Work

Medical Visual Question Answering. Current Med-VQA approaches primarily follow two paradigms: discriminative methods that select from predefined options (Zhang et al., 2023b; Eslami et al., 2023), and generative methods that enable open-ended responses (Bazi et al., 2023; Liu et al., 2023; van Sonsbeek et al., 2023). While discriminative methods achieve high performance in controlled settings, their predefined answer space limits its applicability in real-world medical scenarios. Recent Med-MLLMs (Li et al., 2024b; Jiang et al., 2024c) have shown promising results with flexible response generation. However, they require extensive labeled data for training and fine-tuning. To address this limitation, our AMANDA introduces a novel MLLM agentic framework for data-efficient scenarios without task-specific fine-tuning.

Large Multimodal Agent. Recent research has demonstrated the effectiveness of combining LLMs' reasoning capabilities (OpenAI, 2022, 2023) with MLLMs for visual tasks. Early works like PNP-VQA (Tiong et al., 2022) and Img2LLM (Guo et al., 2023) demonstrated the effectiveness of integrating visual understanding with LLMs' reasoning capabilities. This integration has evolved into sophisticated large multimodal agent systems (You et al., 2023; Surís et al., 2023; Wu et al., 2023c; Xie et al., 2024), where multiple LLM-powered agents collaborate. However, in the medical domain, most existing agent systems (Tang et al., 2023; Fan et al., 2024; Schmidgall et al., 2024; Wei et al., 2024; Li et al., 2024c; Kim et al., 2024) primarily focus on text-based scenarios, lacking crucial multimodal capabilities. While recent work like MMedAgent (Li et al., 2024a) explores multimodal agents for medical applications, it requires extensive task-specific training, limiting its applicability in data-efficient settings. Our AMANDA addresses these limitations by introducing a training-free MLLM agentic framework for data-efficient medical visual reasoning.

Medical Knowledge Augmentation. Integrating medical knowledge has proven essential for enhancing medical AI systems (Fang et al., 2019; Gonzalez-Diaz, 2018; Wang et al., 2020; Chen et al., 2022; Tan et al., 2019; Chen et al., 2020; Soman et al., 2023; Wu et al., 2023a). Representative works include Med-VLP (Chen et al., 2022), which employs UMLS Knowledge Graph (Bodenreider, 2004) for cross-modal alignment, and KG-RAG (Soman et al., 2023), which leverages biomedical knowledge graphs with LLMs. Building upon these advances, our AMANDA introduces a holistic knowledge augmentation approach to enable comprehensive and reliable medical reasoning.

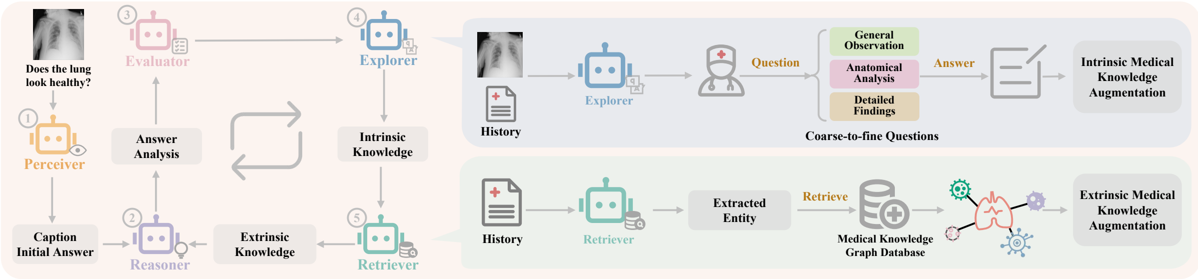


Figure 1: **Overview of our AMANDA framework.** The framework comprises five specialized agents (`Perceiver`, `Reasoner`, `Evaluator`, `Explorer`, and `Retriever`) working collaboratively to enable comprehensive and reliable medical reasoning. Specifically, the `Explorer` incorporates intrinsic medical knowledge through coarse-to-fine question decomposition to enhance reasoning depth, and the `Retriever` integrates extrinsic medical knowledge from biomedical knowledge graphs to enable reliable medical reasoning. The `Evaluator` adaptively controls the depth of Med-KA to enable efficient and accurate medical diagnosis.

3 Proposed Approach – AMANDA

In this section, we first formalize the Med-VQA problem and present our AMANDA framework (Sec. 3.1 and 3.2). We then detail our Med-KA approaches (Sec. 3.3) and present two extensions: the adaptive reasoning refinement mechanism and the few-shot enhancement strategies (Sec. 3.4).

3.1 Problem Formulation

We target Med-VQA in data-efficient scenarios, particularly zero-shot and few-shot settings, where task-specific training data is limited or unavailable. Traditional Med-VQA approaches (Li et al., 2024b; Eslami et al., 2023; Zhang et al., 2023b) typically employ a single Med-MLLM for direct inference. Following previous works (Zhang et al., 2023c), this process can be formulated as:

$$\hat{a} = \Phi_{\text{MedVQA}}\,(\mathcal{I}, q)$$

where $\hat{a}$ is the output answer, $\mathcal{I} \in \mathbb{R}^{H\times W\times C}$ represents the input medical image with height H, width W, and channel number C, q denotes the question, and Φ is the Med-MLLM model.

However, this single-step approach, directly adapted from the general domain, faces two critical limitations in medical image analysis (Liu et al., 2024). First, it fails to systematically examine multiple aspects of medical images, often missing subtle details that are crucial for differentiating similar conditions. Second, in data-efficient scenarios where models encounter novel cases, the lack of comprehensive medical knowledge leads to unreliable analysis or hallucinations (Xia et al., 2024b; Yan et al., 2024).

To address these limitations, we reformulate Med-VQA as an iterative reasoning process that leverages multiple specialized agents:

$$\hat{a}_t = \Phi_{\text{iterative}}(\mathcal{I}, q, \mathcal{H}_{t-1} \cup \bigcup_{i\in\mathcal{A}} h_t^i)$$

where $\hat{a}_t$ represents the refined answer at iteration t, $\Phi_{\text{iterative}}$ denotes our proposed iterative reasoning framework, $\mathcal{H}_{t-1}$ is the accumulated reasoning history up to iteration $t-1$, $\mathcal{A}$ represents our agent set and h_t^i denotes each agent's output at iteration t. This formulation transforms the single-step approach into an iterative reasoning process where specialized agents collaboratively refine the answer through progressive analysis.

3.2 Architecture Overview

To enable such iterative medical reasoning, we design an agentic framework – AMANDA. Our framework comprises three functional modules, where specialized agents work collaboratively:

- **Perception Module.** The `Perceiver` agent, implemented using a Med-MLLM (e.g., LLaVA-Med v1.5 (Li et al., 2024b)), establishes the foundation for visual analysis. Unlike single-step approaches (Li et al., 2024b) that directly generate answers, our `Perceiver` provides two outputs: a detailed medical caption c and an initial answer $\hat{a}_0$ to the main question. The medical caption c is generated through carefully designed prompts (see Appendix H) to systematically describe general observation. The initial answer $\hat{a}_0$, while potentially imperfect, provides a basic foundation that will be progressively refined. Together, these outputs enable more accurate and comprehensive analysis in subsequent modules.
- **Planning Module.** Building upon the Perception Module's outputs, the Planning Module coordi-

nates the overall reasoning process through two LLM-based agents. The Reasoner analyzes the available information (medical caption, initial answer, and any augmented knowledge) to generate a refined answer through systematic medical reasoning. The Evaluator then assesses the reasoning quality through a confidence score, determining whether additional knowledge augmentation is needed (detailed in Sec. 3.4).

- **Action Module.** Triggered by the Planning Module, the Action Module addresses both reasoning bottlenecks through two complementary knowledge augmentation agents. From the intrinsic perspective, the Explorer, powered by LLM, enhances the visual reasoning depth by decomposing the original question q into sub-questions q_{sub}, which are then answered by the same Med-MLLM used in the Perceiver. From the extrinsic perspective, the Retriever, also implemented using LLM, grounds the analysis by retrieving and integrating relevant medical knowledge from biomedical knowledge graphs. Both agents' outputs are fed back to the Planning Module for further answer refinement.

Collaborative Medical Reasoning Workflow. Our AMANDA framework orchestrates these three modules in a collaborative workflow. As shown in Fig. 1: ❶ The Perceiver performs visual analysis to generate a general medical caption and an initial answer. ❷ The Reasoner synthesizes all the available information to produce a refined answer. ❸ The Evaluator assesses the confidence of current answer. ❹ When additional knowledge is needed, the Explorer and Retriever performs both intrinsic Med-KA and extrinsic Med-KA. This augmented knowledge is then fed back to the Reasoner for further refinement.

3.3 Medical Knowledge Augmentation with LLM Agents

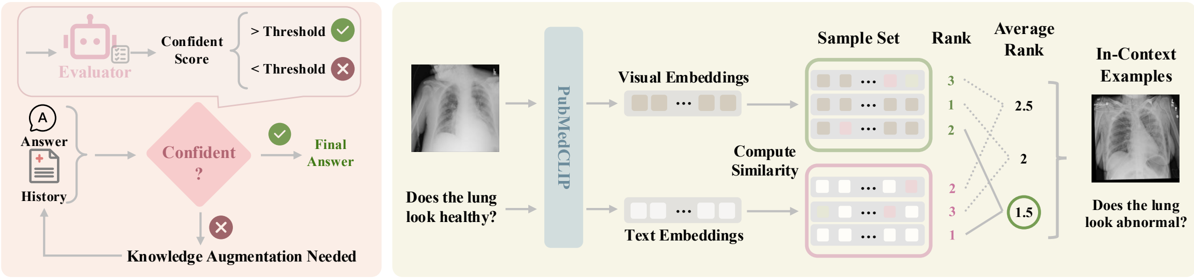
Building upon our agentic framework, we now detail our medical knowledge augmentation strategies that enhance Med-MLLMs' reasoning capability in data-efficient scenarios.

Intrinsic Medical Knowledge Augmentation. In data-efficient scenarios where abundant training data is unavailable, Med-MLLMs often struggle with comprehensive visual analysis due to their single-step inference approach. For instance, when asked *"Does the chest X-ray look healthy?"*, models typically provide general responses like *"no obvious abnormalities"* without examining key diagnostic features. This limitation stems from the lack of progressive questioning in single-step inference, where models fail to focus on specific yet crucial details, resulting in superficial responses that overlook critical diagnostic features.

To address this intrinsic bottleneck, we draw inspiration from the question decomposition strategy, where complex problems are broken down into focused sub-questions for comprehensive analysis. Recent studies have demonstrated that LLMs possess strong capabilities in reasoning enhancement through question decomposition (Wu et al., 2023c; Surís et al., 2023; Zhu et al., 2023; You et al., 2023). These methods leverage LLMs to decompose complex tasks into manageable sub-questions, enabling progressive understanding through structured questioning. Motivated by these advances, we adapt this approach to medical visual analysis to enable deeper and more thorough reasoning.

Specifically, we propose a coarse-to-fine intrinsic Med-KA strategy through our Explorer agent. The strategy is triggered when the Evaluator detects insufficient reasoning depth in the Reasoner's analysis. Our Explorer agent consists of two key components: (1) an LLM-powered questioning component that analyzes the main question, medical caption, and current reasoning history to generate targeted follow-up questions, and (2) an answering component that utilizes the same Med-MLLM as in the Perceiver to provide detailed analysis for each question. At each iteration, Explorer generates three follow-up questions and their corresponding answers in a hierarchical strategy:

- **General Observation.** First focuses on overall appearance and key findings (e.g., *"What is the overall appearance of the image?"*), establishing a foundation for medical analysis.
- **Anatomical Analysis.** Then examines specific anatomical regions or structures, considering their characteristics (size, shape, alignment) and spatial relationships (e.g., *"What is the appearance and position of the cardiac silhouette?"*).
- **Detailed Findings.** Finally investigates potential pathological features in regions of interest (e.g., *"Are there any infiltrates or masses in the lower right lung field, and what are their specific characteristics?"*), enabling the detection of subtle abnormalities through focused analysis.



(a) Adaptive Reasoning Refinement

(b) In-Context Examples Selection

Figure 2: **(a) Adaptive Reasoning Refinement:** The `Evaluator` agent dynamic controls the medical knowledge augmentation process by analyzing the consistency between the current answer and accumulated reasoning history. **(b) In-Context Examples Selection:** The system ranks candidate examples using a dual-similarity metric combining visual and textual features, selecting top-K examples as in-context examples.

This coarse-to-fine approach enhances the intrinsic medical reasoning capability of Med-MLLMs in two ways: (1) breaking down complex analyses into focused steps through hierarchical questioning, enabling thorough examination of diagnostic features; and (2) building a clear reasoning chain that progressively refines visual understanding. Through this progressive analysis, we effectively guide Med-MLLMs to uncover their intrinsic medical knowledge and generate more accurate and detailed diagnostic insights.

Extrinsic Medical Knowledge Augmentation. While our intrinsic Med-KA enhances the depth of medical visual reasoning, Med-MLLMs still face the extrinsic medical reasoning bottleneck due to their static pre-trained knowledge. This issue is particularly critical in data-efficient scenarios where models encounter novel cases that require specialized medical expertise. Without comprehensive domain knowledge, models often generate plausible but incorrect responses, leading to potential hallucinations (Xia et al., 2024b; Yan et al., 2024).

To address this remaining challenge, we introduce an extrinsic Med-KA strategy accomplished by our `Retriever` agent. Inspired by recent advances in Retrieval Augmented Generation (Soman et al., 2024; Xiong et al., 2024), our approach consists of two steps. First, the `Retriever` agent uses an LLM to analyze the accumulated context (including medical captions, questions, and reasoning history) to extract key medical concepts such as "pulmonary nodule". These concepts then serve as queries to SPOKE (Morris et al., 2023), a comprehensive biomedical knowledge graph containing 42 million nodes and 160 million edges assembled from 41 different biomedical databases. Through SPOKE queries, the `Retriever` agent obtains relevant subgraphs containing structured medical knowledge, including disease-symptom associations, anatomical relationships, and medical presentations. These medical facts are then transformed into natural language descriptions for integration into the reasoning process to ground the medical diagnosis.

This extrinsic Med-KA mechanism strengthens Med-MLLMs' reasoning reliability in two ways. First, by retrieving relevant medical knowledge from an external medical knowledge graph, we provide models with specialized expertise needed for novel cases in data-efficient scenarios. Second, the retrieved structured medical facts serve as reliable domain expertise to ground the reasoning process, effectively reducing hallucinations. Together with intrinsic Med-KA, this approach enables Med-MLLMs to perform more reliable medical reasoning through both deeper visual analysis and grounded domain knowledge, especially in data-efficient scenarios.

3.4 Implementation Extensions

Building upon our Med-KA mechanisms, we introduce two extensions to further enhance our framework's effectiveness and efficiency: an adaptive reasoning refinement mechanism, and a few-shot enhancement strategy.

Adaptive Reasoning Refinement. While our two Med-KA mechanisms enhance medical reasoning capabilities, they often require multiple iterations of analysis to achieve comprehensive understanding. However, we observe that excessive refinement can be counterproductive (shown in Fig. 3(a): continuous accumulation of information beyond what's necessary may introduce noise and inconsistencies, potentially overturning initially correct judgments. Moreover, unnecessary iterations increase computational overhead without proportional gains in accuracy. To balance reasoning thoroughness with computational efficiency, we introduce an adaptive reasoning refinement mechanism, implemented through our `Evaluator` agent (Fig.2(a)). The `Evaluator` dynamically controls the knowledge augmentation process by analyzing the consistency between current answers and accumulated reasoning history. It computes a confidence score based on predefined criteria (detailed in AppendixH). When this score exceeds a threshold of 3 out of 5—indicating sufficient reasoning depth and reliability—the system concludes its analysis. If the maximum iteration limit is reached without meeting the confidence threshold, the system adopts the final iteration's response. This adaptive control prevents excessive refinement while ensuring accurate and efficient medical reasoning.

Few-Shot Enhancement. To further demonstrate our framework's effectiveness in data-efficient settings, we extend it to few-shot scenarios via in-context learning. The key challenge lies in selecting the most relevant examples that can effectively guide the reasoning process. To address this, we propose a *dual-similarity selection strategy*. As illustrated in Fig. 2(b), we utilize PubMedCLIP (Zhang et al., 2023b) to compute similarities in both textual and visual domains. Formally, given a test sample with question embedding $\mathcal{T}$ and image embedding $\mathcal{I}$, we select the top K examples from a candidate sample set M through:

$$\mathrm{ICL}_K = \mathrm{TopK}_{i \in M} \frac{1}{2} \left(\mathrm{sim}(\mathcal{T}, \mathcal{T}_i) + \mathrm{sim}(\mathcal{I}, \mathcal{I}_i)\right)$$

where $\mathrm{ICL}_K = \{(c_k, q_k, \hat{a}_k)\}_{k=1}^{K}$ represents the selected examples containing caption, question, and answer triplets. The caption c_k is generated by the `Perceiver` agent from the corresponding medical image. These carefully chosen examples are integrated into our framework, enabling the *Reasoner* to leverage similar cases for more accurate diagnosis. This extension demonstrates our framework's adaptability across both zero-shot and few-shot settings, highlighting its effectiveness in data-efficient medical visual reasoning.

4 Experiments

4.1 Experimental Details

Experimental Setup. We evaluate AMANDA on eight Med-VQA benchmarks that cover diverse medical domains and imaging modalities (detailed in Appendix B). For evaluation models, we primarily use LLaVA-Med-v1.5 (Li et al., 2024b). We also develop variants of Med-InstructBLIP (Dai et al., 2023) and Med-BLIVA (Hu et al., 2024a) both using LLaMA-v1 as their LLM backbone and following LLaVA-Med's training methodology (detailed in Appendix A). Following prior work (Li et al., 2024b), we use accuracy for closed-ended questions and recall for open-ended questions. Additional experiments with general-purpose MLLMs are provided in Appendix D.

Baselines. We compare AMANDA with three types of approaches: (1) Single-step inference by Med-MLLMs serving as our zero-shot baseline; (2) Two-stage methods such as Img2LLM (Guo et al., 2023), which generate image captions via MLLMs before LLM reasoning; and (3) Agent-based approaches like IdealGPT (You et al., 2023) that utilize multiple LLMs for collaborative reasoning.

Implementation Details. Our framework uses GPT-4o as the core reasoning engine for all agents by default. For adaptive reasoning refinement, we set a maximum of 3 iterations and a confidence threshold of 3/5. For few-shot experiments, we use 4 in-context examples as the default setting.

4.2 Effectiveness of AMANDA

Zero-shot Med-VQA. As shown in Table 1 demonstrates the substantial improvements achieved by our framework across different Med-MLLMs and evaluation benchmarks. With LLaVA-Med-v1.5 (Li et al., 2024b), AMANDA achieves an average improvement of **19.36%** over the direct inference baseline. Using Med-BLIVA (Hu et al., 2024a), our method outperforms existing LLM-empowered approaches like Img2LLM (Guo et al., 2023) and IdealGPT (You et al., 2023) by **6.36%** and **5.42%** respectively. These significant improvements stem from our medical-specific design choices. While Img2LLM (Guo et al., 2023) only relies on caption generation and IdealGPT (You et al., 2023) uses general-purpose agent collaboration, our framework enhances medical reasoning through both intrinsic and extrinsic Med-KA along with adaptive reasoning refinement.

Few-shot Med-VQA. We further enhance our framework's effectiveness through few-shot learning, enabling performance gains without model fine-tuning. As shown in Table 1, this few-shot enhancement leads to consistent improvements across

Method	VQA-RAD		SLAKE		IU-Xray	OL3I	OmniMedVQA	FairVL-Med	PMC-OA	Average
	Open	Closed	Open	Closed	Closed	Closed	Closed	Open	Open	
LLaVA-Med-v1.5	30.50	52.94	41.74	44.95	34.50	22.80	40.30	54.58	56.46	42.09
+ Img2LLM	37.81 (+7.31)	47.43 (-5.51)	50.89 (+9.15)	59.86 (+14.91)	70.60 (+36.10)	49.80 (+27.00)	54.40 (+14.10)	61.74 (+7.16)	63.03 (+6.57)	55.06 (+12.97)
+ IdealGPT	41.56 (+11.06)	61.40 (+8.46)	50.96 (+9.22)	69.95 (+25.00)	67.80 (+33.30)	65.40 (+42.60)	53.90 (+13.60)	63.13 (+8.55)	68.02 (+11.56)	60.23 (+18.14)
+ AMANDA	**42.19** (+11.69)	61.03 (+8.09)	54.39 (+12.65)	70.43 (+25.48)	70.30 (+35.80)	65.40 (+42.60)	57.20 (+16.90)	66.60 (+12.02)	65.51 (+9.05)	61.45 (+19.36)
+ AMANDA w/ FS	41.73 (+11.23)	**63.97** (+11.03)	**54.41** (+12.67)	**73.56** (+28.61)	**70.80** (+36.30)	**67.00** (+44.20)	**62.20** (+21.90)	**66.85** (+12.27)	**65.76** (+9.30)	**62.92** (+20.83)
Med-InstructBLIP	32.41	61.76	42.82	59.38	68.60	34.40	29.50	52.18	57.85	48.77
+ Img2LLM	37.61 (+5.20)	57.72 (-4.04)	47.33 (+4.51)	69.23 (+9.85)	73.10 (+4.50)	46.00 (+11.60)	59.60 (+30.10)	59.75 (+7.57)	56.39 (-1.46)	56.30 (+7.53)
+ IdealGPT	40.22 (+7.81)	65.07 (+3.31)	48.85 (+6.03)	65.14 (+5.76)	80.70 (+12.10)	67.40 (+33.00)	56.30 (+26.80)	64.12 (+11.94)	60.10 (+2.25)	60.88 (+12.11)
+ AMANDA	41.02 (+8.61)	68.75 (+6.99)	51.13 (+8.31)	69.47 (+10.09)	79.50 (+10.90)	**67.60** (+33.20)	62.70 (+33.20)	66.61 (+14.43)	63.97 (+6.12)	63.42 (+14.65)
+ AMANDA w/ FS	**46.75** (+14.34)	**74.26** (+12.50)	**52.03** (+9.21)	**72.84** (+13.46)	**84.90** (+16.30)	67.00 (+32.60)	**71.20** (+41.70)	**67.10** (+12.98)	**65.74** (+7.89)	**66.87** (+18.10)
Med-BLIVA	29.19	61.76	43.51	56.01	69.80	38.20	31.90	49.33	54.41	48.24
+ Img2LLM	32.76 (+3.57)	59.93 (-1.83)	44.95 (+1.44)	62.74 (+6.73)	70.10 (+0.30)	46.20 (+8.00)	57.80 (+25.90)	62.43 (+13.10)	55.69 (+1.28)	55.27 (+7.03)
+ IdealGPT	40.84 (+11.65)	53.31 (-8.45)	50.08 (+6.57)	64.66 (+8.65)	71.40 (+1.60)	47.20 (+9.00)	57.80 (+25.90)	64.94 (+15.61)	61.30 (+6.89)	56.84 (+8.60)
+ AMANDA	41.40 (+12.21)	61.76 (+0.00)	**50.95** (+7.44)	68.75 (+12.74)	76.70 (+6.90)	**67.00** (+28.80)	63.20 (+31.30)	66.61 (+17.28)	63.97 (+9.56)	62.26 (+14.02)
+ AMANDA w/ FS	**45.16** (+15.97)	**67.65** (+5.89)	50.49 (+6.98)	**69.23** (+13.22)	**84.60** (+14.80)	65.80 (+27.60)	**65.90** (+34.00)	**67.10** (+17.77)	**65.74** (+11.33)	**64.63** (+16.39)

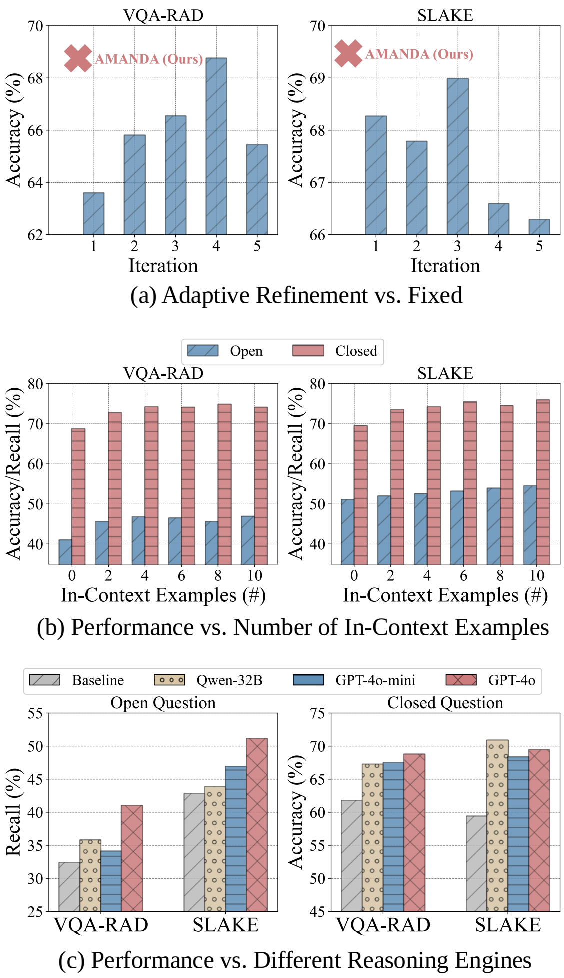
Table 1: **Zero-shot and Few-shot Performance Comparison.** Our framework consistently improves the performance of different Med-MLLMs across various benchmarks. FS denotes experiments with 4 in-context examples.

Model	Hallucination Question Type			Average
	Organ	Condition	Abnormality	
LLaVA-Med-v1.5	39.60	30.30	21.96	30.62
+ AMANDA	88.00 (+48.40)	91.80 (+61.50)	54.00 (+32.04)	77.93 (+47.31)
+ AMANDA w/ FS	**92.40** (+52.80)	**94.80** (+64.50)	**54.40** (+32.44)	**80.53** (+49.91)
Med-InstructBLIP	37.20	16.60	60.60	38.13
+ AMANDA	89.80 (+52.60)	**94.00** (+77.40)	64.40 (+3.80)	82.73 (+44.60)
+ AMANDA w/ FS	**92.00** (+54.80)	93.00 (+76.40)	**65.60** (+5.00)	**83.53** (+45.40)
Med-BLIVA	65.80	53.60	61.80	60.40
+ AMANDA	83.80 (+18.00)	87.80 (+34.20)	61.20 (-0.60)	77.60 (+17.20)
+ AMANDA w/ FS	**90.60** (+24.80)	**92.80** (+39.20)	**64.20** (+2.40)	**82.53** (+22.13)

Table 2: **Effectiveness in reducing hallucination.**

all benchmarks, with Med-InstrcuctBLIP achieving a further **3.45%** gain over its zero-shot performance. These improvements demonstrate the effectiveness of our dual-similarity selection strategy, which provides the `Reasoner` with highly relevant in-context examples to strengthen its medical reasoning capability. These results highlight AMANDA's strong adaptability in data-efficient scenarios, from zero-shot to few-shot settings.

Medical Hallucination Reduction. Beyond improving overall performance, a critical measure of our framework's effectiveness lies in reducing medical hallucinations. We evaluate this capability using ProbMed (Yan et al., 2024), a specialized benchmark for assessing models' medical reasoning reliability. As shown in Table 2, AMANDA achieves substantial reductions in hallucination rates across all tested models, with Med-InstructBLIP (Dai et al., 2023) achieving a **47.37%** reduction. These results demonstrate that our intrinsic and extrinsic Med-KA effectively grounds the medical reasoning process with reliable domain knowledge, addressing a crucial challenge in real-world clinical applications.



(a) Adaptive Refinement vs. Fixed

(b) Performance vs. Number of In-Context Examples

(c) Performance vs. Different Reasoning Engines

Figure 3: Analysis of framework components.

4.3 Further Analysis

Effectiveness of Adaptive Refinement. Fig. 3(a) demonstrates the superiority of our adaptive approach over fixed-iteration strategies. In fixed-iteration settings, performance initially improves with additional iterations but eventually degrades, revealing the detrimental effects of excessive refinement. Our adaptive mechanism achieves dual benefits: it increases accuracy from 66.54% to 68.75% while reducing the average number of iterations from 3.0 to 0.61, resulting in approximately **4.9x** improved efficiency.

Number of In-Context Examples. Fig. 3(b) illustrates how the number of in-context examples affects model performance. While increasing examples initially improves results, the benefits plateau beyond an optimal point. This finding suggests that carefully selected examples are more crucial than quantity for enhancing medical reasoning.

Reasoning Engines Compatibility. As shown in Fig. 3(c), our framework demonstrates compatibility with both closed-source (GPT-4o, GPT-4o-mini) and open-source (DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025)) LLMs as reasoning engines. GPT-4o achieves superior performance on open-ended questions, while open-source alternatives like DeepSeek-R1-Distill-Qwen-32B show competitive results on closed-ended questions. This versatility highlights our method's adaptability across different reasoning engines, enabling users to balance performance requirements with computational cost considerations.

Impact of MLLM Backbones. Table 3 presents a comprehensive analysis of MLLMs with varying backbones and training configurations. Our evaluation reveals three key findings: ❶ larger language backbones generally achieve better performance, particularly on closed-ended questions where precise reasoning is crucial; ❷ increasing the pre-training dataset size from 60K (Li et al., 2024b) to 150K (Cui et al., 2024) samples leads to significant improvements across all metrics; and ❸ models with medical domain pre-training like PMC-LLaMA (Wu et al., 2023b) demonstrate strong performance, highlighting the value of domain-specific knowledge in medical reasoning.

4.4 Ablation Study

We conduct systematic ablation experiments to evaluate each agent's contribution to our framework. ❶ Removing `Perceiver` eliminates the foundation for understanding query images, resulting in significant performance degradation. ❷ Without `Explorer`, the framework loses its ability to progressively uncover key diagnostic features, limiting the depth of medical reasoning. ❸ The absence of `Retriver` reduces performance by removing access to extrinsic domain expertise. ❹ Without `Reasoner`, the framework cannot effectively analyze accumulated information and refine answers, leading to lower accuracy. ❺ The `Evaluator` agent proves crucial for efficiency: while open-ended questions benefit from extended reasoning cycles, closed-ended questions suffer from unnecessary refinements that can introduce noise and contradictions. Moreover, the `Evaluator` substantially reduces the average number of iterations while maintaining performance. These results collectively validate each agent's essential role in achieving efficient and accurate medical reasoning.

Model	Model Size	Dataset Size	VQA-RAD		SLAKE	
			Open	Closed	Open	Closed
LLaMA	7B	60K	41.40	61.76	50.95	68.75
LLaMA	13B	60K	38.34	**66.54**	<u>51.85</u>	**69.47**
LLaMA	7B	150K	**47.90**	<u>66.18</u>	51.25	68.27
Vicuna	7B	60K	<u>41.63</u>	58.82	**51.90**	67.31
PMC-LLaMA	7B	60K	40.80	62.87	51.01	<u>68.75</u>

Table 3: **Analysis of language backbones in MedBLIVA.** Each column's highest score is in **bold**, while the second highest score is <u>underlined</u>.

Method	VQA-RAD		SLAKE	
	Open	Closed	Open	Closed
AMANDA	42.19	61.03	54.39	70.43
- Perceiver	22.70 (-19.49)	40.81 (-20.22)	28.72 (-25.67)	35.58 (-34.85)
- Explorer	38.82 (-3.37)	56.62 (-4.41)	50.28 (-4.11)	64.66 (-5.77)
- Retriever	41.11 (-1.08)	60.29 (-0.74)	52.90 (-1.49)	69.47 (-0.96)
- Reasoner	38.09 (-4.10)	57.72 (-3.31)	50.21 (-4.18)	68.03 (-2.40)
- Evaluator	43.56 (+1.37)	57.35 (-3.68)	54.72 (+0.33)	69.23 (-1.20)

Table 4: **Ablation study.** Analysis of different agents by removing each from the full model.

5 Conclusion

In this work, we present AMANDA, a training-free agentic framework that addresses Med-MLLMs' intrinsic and extrinsic bottlenecks in data-efficient scenarios. Our framework enhances medical visual reasoning through coarse-to-fine question decomposition and grounds its analysis with extrinsic knowledge graphs, while maintaining efficiency through adaptive reasoning refinement. Extensive experiments demonstrate substantial improvements on Med-VQA in both zero-shot and few-shot settings, highlighting AMANDA's potential for reliable AI-assisted medical diagnosis in resource-constrained environments.

6 Limitations

While our work demonstrates promising results, several perspectives remain for future exploration. First, although we evaluate on eight diverse Med-VQA benchmarks, testing on more specialized medical datasets across different modalities (e.g., MRI, CT) could further validate our framework's generalizability. Second, our experiments primarily focus on publicly available Med-MLLMs with language models up to 13B parameters; investigating the impact of larger language models (e.g., 70B) could potentially reveal additional performance gains. Third, incorporating more diverse external medical knowledge resources (e.g., medical textbooks, clinical guidelines, and medical reports) could potentially enhance our framework's capability in handling various types of medical queries. Fourth, enabling our agents to utilize existing medical tools and collaborate with hospitals for diagnosis would be a promising direction for real-world deployment. Finally, while we focus on a training-free approach, exploring lightweight fine-tuning strategies could potentially achieve better performance improvements while maintaining reasonable computational requirements in resource-constrained scenarios.

References

Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Laila Bashmal, and Mansour Zuair. 2023. Vision–language model for visual question answering in medical imagery. *Bioengineering*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*.

Zhihong Chen, Guanbin Li, and Xiang Wan. 2022. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *Proceedings of the 30th ACM International Conference on Multimedia*.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.

Hejie Cui, Lingjun Mao, Xin Liang, Jieyu Zhang, Hui Ren, Quanzheng Li, Xiang Li, and Carl Yang. 2024. Biomedical visual instruction tuning with clinician preference alignment. *arXiv preprint arXiv:2406.13173*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*.

Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. 2023. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Interactive evaluation and collaboration of llms as intern doctors for clinical diagnosis. *arXiv preprint arXiv:2402.09742*.

Leyuan Fang, Chong Wang, Shutao Li, Hossein Rabbani, Xiangdong Chen, and Zhimin Liu. 2019. Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification. *IEEE transactions on medical imaging*.

Ivan Gonzalez-Diaz. 2018. Dermaknet: Incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis. *IEEE journal of biomedical and health informatics*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. 2023. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Iryna Hartsock and Ghulam Rasool. 2024. Vision-language models for medical report generation and visual question answering: A review. *arXiv preprint arXiv:2403.02469*.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. 2024a. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024b. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Bowen Jiang, Yangxinyu Xie, Xiaomeng Wang, Weijie J Su, Camillo J Taylor, and Tanwi Mallick. 2024a.

Multi-modal and multi-agent systems meet rationality: A survey. *arXiv preprint arXiv:2406.00252*.

Bowen Jiang, Zhijun Zhuang, Shreyas S Shivakumar, Dan Roth, and Camillo J Taylor. 2024b. Multi-agent vqa: Exploring multi-agent foundation models in zero-shot visual question answering. *arXiv preprint arXiv:2403.14783*.

Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. 2024c. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. *Preprint*, arXiv:2404.10237.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Adaptive collaboration strategy for llms in medical decision making. *arXiv preprint arXiv:2404.15155*.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*.

Binxu Li, Tiankai Yan, Yuanting Pan, Zhe Xu, Jie Luo, Ruiyang Ji, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. 2024a. Mmedagent: Learning to use medical tools with multi-modal agent. *arXiv preprint arXiv:2407.02483*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024b. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024c. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer.

Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. 2023b. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *International Symposium on Biomedical Imaging (ISBI)*. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*.

Yunyi Liu, Zhanyu Wang, Dong Xu, and Luping Zhou. 2023. Q2atransformer: Improving medical vqa via an answer querying decoder. *arXiv preprint arXiv:2304.01611*.

Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, et al. 2024. Fairclip: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

John H Morris, Karthik Soman, Rabia E Akbas, Xiaoyuan Zhou, Brett Smith, Elaine C Meng, Conrad C Huang, Gabriel Cerono, Gundolf Schenk, Angela Rizk-Jackson, et al. 2023. The scalable precision medicine open knowledge engine (spoke): a massive knowledge graph of biomedical information. *Bioinformatics*.

OpenAI. 2022. ChatGPT. `https://openai.com/blog/chatgpt/`.

OpenAI. 2023. GPT-4 technical report. `https://arxiv.org/abs/2303.08774`. *Preprint*, arXiv:2303.08774.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, et al. 2023. Biomedical knowledge graph-enhanced prompt generation for large language models. *arXiv preprint arXiv:2311.17330*.

Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, et al. 2024. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*.

Jiaxing Tan, Yumei Huo, Zhengrong Liang, and Lihong Li. 2019. Expert knowledge-infused deep learning for automatic lung nodule detection. *Journal of X-ray Science and Technology*.

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

Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. 2022. Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. *arXiv preprint arXiv:2210.08773*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tom van Sonsbeek, Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees GM Snoek, and Marcel Worring. 2023. Open-ended medical visual question answering through prefix tuning of language models. *arXiv preprint arXiv:2303.05977*.

Kun Wang, Xiaohong Zhang, Sheng Huang, Feiyu Chen, Xiangbo Zhang, and Luwen Huangfu. 2020. Learning to recognize thoracic disease in chest x-rays with knowledge-guided deep zoom neural networks. *IEEE Access*.

Zeqing Wang, Wentao Wan, Runmeng Chen, Qiqing Lao, Minjie Lang, and Keze Wang. 2023. Towards top-down reasoning: An explainable multi-agent approach for visual question answering. *arXiv preprint arXiv:2311.17331*.

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

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023b. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023c. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. 2024a. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *arXiv preprint arXiv:2406.06007*.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. 2024b. Rule: Reliable multimodal rag for factuality in medical vision language models. *arXiv preprint arXiv:2407.05131*.

Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. 2024. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. 2024. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*.

Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. 2024. Worse than random? an embarrassingly simple probing evaluation of large multimodal models in medical vqa. *arXiv preprint arXiv:2405.20421*.

Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad A Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. 2023. Idealgpt: Iteratively decomposing vision and language reasoning via large language models. *arXiv preprint arXiv:2305.14985*.

Juan M Zambrano Chaves, Andrew L Wentland, Arjun D Desai, Imon Banerjee, Gurkiran Kaur, Ramon Correa, Robert D Boutin, David J Maron, Fatima Rodriguez, Alexander T Sandhu, et al. 2023. Opportunistic assessment of ischemic heart disease risk using abdominopelvic computed tomography and medical record data: a multimodal explainable artificial intelligence approach. *Scientific reports*.

Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, Lifang He, Brian D. Davison, Quanzheng Li, Yong Chen, Hongfang Liu, and Lichao Sun. 2023a. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *Preprint*, arXiv:2305.17100.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. 2023b. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023c. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. 2023. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594*.

A Details of Evaluated MLLMs

We evaluate our framework across both medical domain-specific and general-domain MLLMs to demonstrate its versatility and effectiveness.

A.1 Medical Domain-Specific MLLMs

- **LLaVA-Med-v1.5**(Li et al., 2024b): Built on Mistral-7B(Jiang et al., 2023), this is our primary evaluation model. It extends LLaVA (Liu et al., 2024) for medical domain understanding through specialized training on medical image-text pairs and conversational data.
- **Med-InstructBLIP**: Our medical adaptation of InstructBLIP (Dai et al., 2023) using LLaMa-7B (Touvron et al., 2023). Following LLaVA-Med's training methodology (Li et al., 2024b), we adapt the model for medical visual understanding while maintaining its instruction-tuning capabilities.
- **Med-BLIVA**: A medical version of BLIVA (Hu et al., 2024a) based on LLaMa-7B (Touvron et al., 2023). We adapt it using LLaVA-Med's training strategy (Li et al., 2024b) to combine BLIVA's visual reasoning capabilities with medical domain expertise.

A.2 Pre-training Details of Med-MLLMs

For Med-InstructBLIP and Med-BLIVA, we follow LLaVA-Med's (Li et al., 2024b) two-stage training strategy:

- **Stage 1: Feature Alignment.** We first align the visual features with medical concepts through projection learning. Using 600K filtered image-text pairs from PMC-15M, we train only the projection layer while keeping both the visual encoder and language model frozen. This stage enables the models to understand biomedical visual concepts efficiently.
- **Stage 2: Instruction Tuning.** We then perform end-to-end instruction tuning with the projection layer and language model unfrozen. Using 60K medical image-text instruction data, we train the models to follow various medical instructions and perform visual reasoning tasks. This stage enhances the models' capabilities in medical visual understanding and dialogue interaction.

A.3 General-Domain MLLMs

- **InstructBLIP** (Dai et al., 2023): A strong general-domain MLLM with instruction-tuning capabilities. We evaluate it using its original pre-trained weights to assess our framework's effectiveness on models without medical domain adaptation.
- **xGen-MM** (Xue et al., 2024): The latest BLIP architecture variant with advanced visual reasoning capabilities. We use its original weights to test our framework's compatibility with state-of-the-art general-purpose MLLMs.

Evaluating these general-domain models alongside medical-specific ones demonstrates our framework's versatility across different architectures and its ability to enhance medical reasoning capabilities regardless of domain specialization.

B Details of Med-VQA Benchmarks

We utilize open-source Med-VQA benchmarks, which cover a wide range of medical image modalities and anatomical regions: VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), IU-Xray (Demner-Fushman et al., 2016), Harvard-FairVLMed (Luo et al., 2024), PMC-OA (Lin et al., 2023a), OL3I (Zambrano Chaves et al., 2023), OmniMedVQA (Hu et al., 2024b), and ProbMed (Yan et al., 2024). Table 5 provides comprehensive statistics about these datasets. The details of each benchmark are as follows:

- **VQA-RAD** (Lau et al., 2018): A dedicated Med-VQA dataset containing 315 medical images and 3,515 question-answer pairs. It covers various medical imaging modalities including chest X-rays and CT scans. The questions are carefully designed to evaluate both visual understanding and clinical reasoning capabilities, categorized into different types including modality, plane, organ system, and abnormality detection.
- **SLAKE** (Liu et al., 2021): A comprehensive Med-VQA dataset comprising 14,028 question-answer pairs on 8,851 medical images across multiple modalities (CT, MRI, X-Ray). The questions assess different levels of understanding, from basic pattern recognition to complex clinical reasoning. The dataset contains 11,222 training samples and 1,061 testing samples.
- **IU-Xray** (Demner-Fushman et al., 2016): A spe-

Table 5: **Comprehensive statistics of the Med-VQA Benchmarks.**

Index	Data Source	Modality	Region	# Images	# QA Items	Answer Type	# Test
1	VQA-RAD (Lau et al., 2018)	X-Ray, CT	Chest, Abd	315	3,515	Mixed	451
2	SLAKE (Liu et al., 2021)	CT, MRI, X-Ray	Mixture	8,851	14,028	Open-ended	1,061
3	IU-Xray (Demner-Fushman et al., 2016)	X-Ray	Chest	589	2,573	Yes/No	1,000
4	Harvard-FairVLMed (Luo et al., 2024)	Fundus	Eye	713	2,838	Open-ended	1,000
5	OL3I (Zambrano Chaves et al., 2023)	CT	Heart	1,000	1,000	Yes/No	500
6	PMC-OA (Zhang et al., 2023c)	Mixture	Mixture	2,587	13,294	Open-ended	1,000
7	OmniMedVQA (Hu et al., 2024b)	Mixture*	Mixture	10,995	12,227	Multi-choice	1,000
8	ProbMed (Yan et al., 2024)	Mixture*	Mixture	6,303	57,132	Yes/No	1,500

cialized dataset focusing on chest X-ray images and their corresponding diagnostic reports. Our benchmark includes 589 frontal chest X-rays from the test set, along with their detailed clinical reports.

- **Harvard-FairVLMed** (Luo et al., 2024): A multimodal dataset of fundus images designed to evaluate fairness in AI models. It contains image and text data from diverse demographic groups, specifically focusing on bias assessment in medical visual understanding.

- **PMC-OA** (Lin et al., 2023a): A large-scale collection of biomedical images extracted from open-access publications. We incorporate 2,587 diverse image-text pairs randomly selected from the test set into our benchmark.

- **OL3I** (Zambrano Chaves et al., 2023): A publicly available dataset focused on predicting ischemic heart disease (IHD) using contrast-enhanced abdominal-pelvic CT examinations. It features a retrospective cohort with up to 5 years of follow-up data.

- **OmniMedVQA** (Hu et al., 2024b): A comprehensive Med-VQA benchmark collected from 73 different medical datasets. It encompasses 12 different imaging modalities and covers more than 20 distinct anatomical areas, providing broad coverage of medical visual understanding tasks.

- **ProbMed** (Yan et al., 2024): A specialized benchmark designed for evaluating model hallucination, comprising 6,303 images and 57,132 question-answer pairs. It includes carefully designed adversarial QA pairs across three modalities (X-ray, MRI, CT scan) and four anatomical regions (abdomen, brain, chest, spine).

B.1 Evaluation Protocol

Following (Xia et al., 2024a), we construct our evaluation benchmark using diverse medical image-text pairs from eight datasets. For classic Med-VQA benchmarks VQA-RAD and SLAKE, we use their complete test sets (451 and 1,061 QA pairs respectively) to maintain consistency with previous works. For larger-scale datasets (IU-Xray, Harvard-FairVLMed, OL3I, PMC-OA, OmniMedVQA, and ProbMed), we randomly sample 500-1,500 test examples from their original test sets due to computational constraints.

The remaining training samples from these datasets serve as our in-context learning pool for few-shot evaluation. For each test image, we retrieve similar examples based on visual and semantic similarity to construct few-shot prompts. This diverse collection of datasets, covering various modalities and answer formats (Yes/No, Open-ended, and Multi-choice), enables comprehensive evaluation of medical visual understanding capabilities.

C Evaluation Metrics

For the closed-ended questions, we report the accuracy in a more strict way compared to prior work (Li et al., 2024b). Instead of checking whether the ground-truth answer appears anywhere in the generated response, we only consider the first occurring yes/no-type word as the final prediction. This eliminates the inflated accuracy caused by long generated texts that include both "yes" and "no". For open-ended questions, we use recall to evaluate the ratio of ground-truth tokens that appear in the generated sequences. Different from the literature that selects from a fixed set of training answers, we do not provide any constraints on the model's open-ended responses. This makes our formulation closer to real open-ended questions but is intrinsically more challenging. For a fair compari-

Method	VQA-RAD		SLAKE		IU-Xray	OL3I	OmniMedVQA	FairVL-Med	PMC-OA	Average
	Open	Closed	Open	Closed	Closed	Closed	Closed	Open	Open	
General MLLMs (without Medical Pre-training)										
InstructBLIP	16.09	62.50	22.14	59.86	62.30	36.11	33.40	45.22	42.90	42.28
+ AMANDA	29.86 (+13.77)	65.81 (+3.31)	41.03 (+18.89)	66.35 (+6.49)	68.30 (+6.00)	61.11 (+25.00)	52.30 (+18.90)	**64.83** (+19.61)	63.08 (+20.18)	56.96 (+14.68)
+ AMANDA w/ FS	**38.96** (+22.87)	**68.01** (+5.51)	**48.61** (+26.47)	**69.71** (+9.85)	**71.30** (+9.00)	**63.89** (+27.78)	**54.40** (+21.00)	64.81 (+19.59)	**63.12** (+20.22)	**60.31** (+18.03)
Xgen-MM	16.08	62.50	22.14	59.86	53.30	37.80	44.70	58.38	49.19	44.88
+ AMANDA	35.20 (+19.12)	67.28 (+4.78)	46.47 (+24.33)	70.19 (+10.33)	59.20 (+5.90)	48.80 (+11.00)	54.10 (+9.40)	67.34 (+8.96)	**64.85** (+15.66)	57.05 (+12.17)
+ AMANDA w/ FS	**37.76** (+21.68)	**75.37** (+12.87)	**47.92** (+25.78)	**74.28** (+14.42)	**69.60** (+16.30)	**51.60** (+13.80)	**58.10** (+13.40)	**67.42** (+9.04)	64.72 (+15.53)	**60.75** (+15.87)

Table 6: **Generalization to general-purpose MLLMs.** Zero-shot and few-shot results across Med-VQA benchmarks using general MLLMs, showing the framework's strong generalization capability beyond Med-MLLMs.

LLM Engine	Method	VQA-RAD		SLAKE	
		Open	Closed	Open	Closed
DeepSeek-R1-Distill-Qwen-32B	**Med-InstructBLIP**	32.41	61.76	42.82	59.38
	+ AMANDA	35.81 (+3.40)	67.28 (+5.52)	43.87 (+1.05)	70.91 (+11.53)
DeepSeek-R1-Distill-Llama-70B	**Med-InstructBLIP**	32.41	61.76	42.82	59.38
	+ AMANDA	34.28 (+1.87)	66.18 (+4.42)	44.34 (+1.52)	70.43 (+11.05)

Table 7: Performance of different open source LLMs as reasoning engine on VQA-RAD and SLAKE datasets.

Method	VQA-RAD		SLAKE	
	Open	Closed	Open	Closed
SIRI (Wang et al., 2023)	-	45.80	-	-
KG-RAG (Soman et al., 2024)	35.56	52.57	46.71	66.34
BiomedGPT-S (Zhang et al., 2023a)	13.40	57.80	66.50	73.40
AMANDA	42.19	61.03	54.39	70.43

Table 8: Comparison of different methods on VQA-RAD and SLAKE datasets.

Metric	VQA-RAD		SLAKE	
	Open	Closed	Open	Closed
Average	42.80	61.32	54.12	70.28
Std	0.79	0.88	0.82	0.47
CV	0.02	0.01	0.02	0.01

Table 9: Stability analysis of AMANDA across 5 runs with different random seeds. Std represents standard error and CV denotes coefficient of variation

son, we use the same strict accuracy metric for all methods. While this might lead to lower absolute numbers compared to what is typically reported, we believe it better reflects the true performance and is more meaningful.

D Additional Results of AMANDA Framework on General MLLMs

While our main experiments demonstrate the effectiveness of AMANDA on medical-specialized MLLMs, we further evaluate its generalization capability on general-domain MLLMs that lack medical pre-training. As shown in Table 6, our framework demonstrates strong generalization capability across different models. Specifically, when applied to InstructBLIP (Dai et al., 2023), AMANDA achieves an average improvement of **14.68%** over direct inference. These results suggest that our framework can effectively bridge the domain gap and enable general-purpose MLLMs to perform reliable medical visual reasoning.

E Compatibility with Different LLM Engines

To demonstrate the versatility of AMANDA, we evaluate its performance using different open-source LLMs as reasoning engines. As shown in Table 7, we test our framework with DeepSeek-R1-Distill-Qwen-32B and DeepSeek-R1-Distill-Llama-70B (Guo et al., 2025) on the VQA-RAD and SLAKE datasets. When integrated with Med-InstructBLIP, both models show substantial improvements across all question types. Notably, with DeepSeek-R1-Distill-Qwen-32B, we achieve significant gains on closed-ended questions (+5.52% on VQA-RAD, +11.53% on SLAKE), while maintaining competitive performance on open-ended questions. Similar improvements are observed with DeepSeek-R1-Distill-Llama-70B, demonstrating that AMANDA can effectively enhance medical visual reasoning capabilities regardless of the underlying LLM engine. These results indicate that our framework provides a cost-effective solution for improving Med-VQA performance without requiring specialized training or extensive computational resources.

E.1 Comparison with Strong Baselines

To provide a more comprehensive evaluation, we compare AMANDA with several strong baselines, including both zero-shot and supervised approaches. The results in Table 8 demonstrate AMANDA's effectiveness across different evaluation settings. Our framework significantly outperforms other zero-shot approaches, including SIRI (Wang et al., 2023) (a multi-agent framework) and KG-RAG (Soman et al., 2024) (which combines knowledge retrieval with LLM reasoning). Notably, AMANDA achieves superior performance on VQA-RAD and competitive results on SLAKE compared to BiomedGPT-S (Zhang et al., 2023a), despite the latter's advantage of supervised training on downstream tasks. These comprehensive comparisons validate the effectiveness of our training-free approach in medical visual reasoning tasks.

E.2 Framework Stability

We have thoroughly evaluated our framework's stability. As shown in Table 9, we have conducted additional experiments, running LLaVA-Med v1.5 with AMANDA 5 times with different seeds on different benchmarks. These results demonstrate the high stability of our framework, with standard deviations consistently below 1% and coefficients of variation as low as 0.01-0.02. To put these variations in perspective, they are significantly smaller than the performance improvements our framework achieves over the baseline (e.g., an 8-25% absolute improvement), confirming that AMANDA provides stable and reliable enhancements across different medical visual reasoning tasks and models.

F Pseudo-Code of AMANDA Framework

The algorithm illustrates how our framework orchestrates multiple specialized agents for collaborative medical reasoning. The process operates as follows:

- The `Perceiver` agent first analyzes the medical image and generates a detailed caption along with an initial answer, establishing a foundation for visual understanding.
- The `Reasoner` agent then processes this initial information to generate a preliminary medical analysis based on the visual findings.
- The `Evaluator` agent assesses the confidence of the current answer by analyzing its consistency with the accumulated evidence.

Algorithm 1 AMANDA Framework Pipeline

```
def AMANDA(I: Image, Q: str) -> str:
    """
    Data-efficient Med-VQA
    Args:
        I: Input medical image
        Q: Input question
    Returns:
        Final answer
    """
    # Initialize reasoning history
    H = []

    # Initial Visual Understanding
    C, A_0 = Perceiver(I, Q)  # Generate
        medical caption and initial
        answer
    H.append((C, A_0))
    A_0 = Reasoner(Q, H)  # Initial
        reasoning
    confidence = Evaluator(A_0, H)

    # Medical Knowledge Augmentation
    while confidence < THRESHOLD:
        # Intrinsic Med-KA
        Q_sub, A_sub = Explorer(Q, H)
        H.append((Q_sub, A_sub))

        # Extrinsic Med-KA
        K = Retriever(H)
        H.append(K)

        # Re-reasoning with Enhanced
            Knowledge
        A_t = Reasoner(Q, H)
        confidence = Evaluator(A_t, H)

    return A_t
```

- When confidence is insufficient, the `Explorer` agent generates strategic follow-up questions to probe deeper into critical visual details, while the *Retriever* agent supplements the analysis with relevant medical knowledge from external sources.
- This iterative process continues until the `Evaluator` determines that sufficient confidence has been achieved, ensuring both comprehensive analysis and reliable diagnosis.

G Case Study

As shown in Table 10, this case study demonstrates how our AMANDA framework effectively corrects initial misdiagnosis through comprehensive medical knowledge augmentation. Initially, the Med-MLLM baseline incorrectly identifies a rightward mediastinal shift. Our framework then initiates a systematic analysis through three key components. First, the `Perceiver` generates a detailed medical caption, establishing a foundation for understanding the image's key features. Second, through intrinsic Med-KA, the `Explorer` generates strategically designed sub-questions that progressively examine the mediastinal position from different perspectives. Third, the `Retriever` provides relevant medical knowledge about the mediastinum. Based on this augmented knowledge, the `Reasoner` generates a refined answer, which the `Evaluator` determines to be highly confident, demonstrating the effectiveness of our adaptive reasoning refinement. This case study exemplifies how our framework enhances medical reasoning through structured analysis and knowledge augmentation, leading to efficient and accurate diagnosis.

H Prompt Design of AMANDA

In this section, we provide detailed prompt designs for each agent in our framework. First, Table 11 presents the caption generation prompts used by the `Perceiver` agent for comprehensive medical image understanding. Then, Table 12 shows the structured instructions for the `Explorer` agent to generate coarse-to-fine follow-up questions. Table 15 details how the `Evaluator` agent assesses reasoning confidence. Finally, Tables 14 and 13 provide the `Reasoner` agent's prompts for handling closed-ended and open-ended questions respectively, demonstrating our framework's ability to adapt to different question types.

Table 10: A case study AMANDA showing how our framework progressively refines the answer through knowledge augmentation.

Main Question: Has the midline of the mediastinum shifted?

Medical Caption:
This is a chest X-ray, which is a type of medical imaging study that uses low-energy X-rays to visualize the structures within the chest, including the lungs, heart, and bones. The chest X-ray shows bilateral ground-glass opacities in both lungs, which can be indicative of various lung conditions such as infections, inflammation, or other abnormalities.

Initial Answer: Yes, the midline of the mediastinum has shifted to the right.
Evaluator: Confident Score : 1

Coarse-to-fine question decomposition:
Q1: Are there any visible signs of mediastinal shift, such as displacement of the trachea or heart?
A1: No, the chest X-ray does not show any visible signs of mediastinal shift, such as displacement of the trachea or heart.

Q2: Is the position of the heart and trachea symmetrical and centered within the thoracic cavity?
A2: Yes, the position of the heart and trachea appears to be symmetrical and centered within the thoracic cavity.

Q3: Are there any abnormalities in the lung volumes or pleural spaces that could contribute to a shift in the mediastinum?
A3: No, there are no abnormalities in the lung volumes or pleural spaces that could contribute to a shift in the mediastinum.

Retrieved Medical Knowledge:
Lung disease Localizes in: cavity, chest, diaphragm, mediastinum.

Iteration 1 Answer: No, the midline of the mediastinum has not shifted.
Evaluator: Confident Score : 4

Final Answer: No, the midline of the mediastinum has not shifted.
Ground Truth: No.

Table 11: Prompts for the `Perceiver` Agent. Following LLaVA-Med's instruction tuning strategy, we randomly select one of these prompts to guide the `Perceiver` agent in generating comprehensive medical image descriptions.

PERCEIVER_CAPTION_PROMPTS:

- Describe the following image in detail
- Provide a detailed description of the given image
- Give an elaborate explanation of the image you see
- Share a comprehensive rundown of the presented image
- Offer a thorough analysis of the image
- Explain the various aspects of the image before you
- Clarify the contents of the displayed image with great detail
- Characterize the image using a well-detailed description
- Break down the elements of the image in a detailed manner
- Walk through the important details of the image
- Portray the image with a rich, descriptive narrative
- Narrate the contents of the image with precision
- Analyze the image in a comprehensive and detailed manner
- Illustrate the image through a descriptive explanation
- Examine the image closely and share its details
- Write an exhaustive depiction of the given image

Table 12: `Explorer` agent instructions for generating follow-up questions.

EXPLORER_SYSTEM_PROMPT:

You are an AI language model tasked with helping clinicians analyze medical images. Your goal is to decompose a primary clinical question into several sub-questions. By answering these sub-questions, it will be easier to arrive at a comprehensive answer for the main question.

Instruction: Given a general caption that might not be entirely precise but provides an overall description, and a clinical question, generate a series of sub-questions to help thoroughly answer the main question. These sub-questions should guide the analysis step by step, focusing on the different aspects that could influence the final answer, keeping in mind clinical relevance and imaging characteristics.

Rules:

- Break down the question into smaller parts following this hierarchical approach:
  (a) First, ask about general/overall observations
  (b) Then, focus on specific anatomical regions or structures
  (c) Finally, ask about detailed findings or specific characteristics
- Consider these aspects in your questions:
  - Presence or absence of specific findings
  - Characteristics of structures (e.g., size, shape, alignment)
  - Orientation and positioning of the patient or organs
  - Comparison of abnormal vs. normal findings
- The number of sub-questions should be less or equal to {max_sub_questions}.
- Order your questions from general to specific (coarse to fine-grained).

Format:

Sub-question 1: [General observation question]
Sub-question 2: [Specific anatomical region question]
Sub-question 3: [Detailed finding question]
...

EXPLORER_PROMPT:

Image description: {caption}
Main question: {question}
History: {history}
Please generate a series of follow-up questions following a coarse-to-fine approach. Start with general observations and progressively move to more specific details.

Table 13: Open-ended `Reasoner` instructions.

OPEN_ENDED_REASONER_SYSTEM_PROMPT:

You are a medical AI assistant with rich visual commonsense knowledge and strong reasoning abilities.

You will be provided with:

1. A main question about an image.
2. An imperfect initial answer to the main question provided by a visual AI model. Note that the answers may not be entirely precise.
3. A general caption that might not be entirely precise but provides an overall description.
4. Some conversation history containing follow-up questions and answers.
5. Some grounded medical information.
6. Some similar examples with their answers for reference.

Your goal: Based on the above information, find the answer to the main question.

Rules:

1. Begin with a brief paragraph demonstrating your reasoning and inference process. Start with the format: "Analysis:".
2. Be logical and consistent in evaluating all clues, including as many relevant details as possible.
3. Use similar examples to inform your reasoning.

Response Format:

```
Analysis: xxxxxx.

Answer: xxxxxx
```

OPEN_ENDED_REASONER_PROMPT:

Imperfect image description: {caption}
Open-ended question: {question}
Initial answer: {initial_answer}
History:
{history}
Additional information: {rag_context}
Please provide a detailed answer to the open-ended question based on all the information provided.

Table 14: Closed-ended `Reasoner` instructions.

CLOSED_ENDED_REASONER_SYSTEM_PROMPT:

You are a medical AI assistant with rich visual commonsense knowledge and strong reasoning abilities.

You will be provided with:

1. A main question about an image.
2. An imperfect initial answer to the main question provided by a visual AI model. Note that the answers may not be entirely precise.
3. A general caption that might not be entirely precise but provides an overall description.
4. Some conversation history containing follow-up questions and answers.
5. Some grounded medical information.
6. Some similar examples with their answers for reference.

Your goal: Based on the above information, find the answer to the main question.

Rules:

1. Begin with a brief paragraph demonstrating your reasoning and inference process. Start with the format: "Analysis:".
2. Be logical and consistent in evaluating all clues, but aim to preserve the initial answer unless strong contradictions arise.
3. Use similar examples to inform your reasoning.

Response Format:

```
Analysis: xxxxxx.

Answer: [Yes/No] or [Selected Option]
```

CLOSED_ENDED_REASONER_PROMPT:

Imperfect image description: {caption}
Closed-ended question: {question}
Initial answer: {initial_answer}
History:
{history}
Additional information: {rag_context}
Please provide an answer to the closed-ended question based on all the information provided.

Table 15: `Evaluator` agent instructions for assessing confidence levels in medical image analysis responses.

EVALUATOR_SYSTEM_PROMPT:

You are a medical AI assistant specialized in evaluating answers for medical image analysis.
You will be provided with:

1. A main question about a medical image.
2. A general caption that might not be entirely precise and may contain false information.
3. Current answer.
4. History of the conversation.
5. Examples from in-context learning.

Your goal:

1. Assess the confidence level of a given answer and provide a brief explanation.
2. Provide a confidence score from 1 to 5, where 1 means completely uncertain and 5 means very certain.
3. Use examples from in-context learning to assist in evaluating the answer.

Evaluation Criteria:

- **Contradictory Evidence**: Look for any information that strongly contradicts the current answer. If significant conflicting information is found, reduce the confidence level.

Scoring Guidance:

- **Score 5**: The answer is accurate, consistent with all provided information, and has no significant conflicting evidence.
- **Score 4**: The answer is mostly correct, with minor issues or slight uncertainty.
- **Score 3**: The answer is generally acceptable, with some uncertainty or minor inconsistencies, but it mostly aligns with the question.
- **Score 2**: The answer has notable inaccuracies or lacks consistency, with some conflicting information present.
- **Score 1**: The answer is largely incorrect, inconsistent, or contains major contradictions with the provided information.

Response Format:

```
Score: [1-5]
Explanation: [Your explanation]
```

EVALUATOR_PROMPT:

Imperfect image description: {caption}
Main question: {question}
Current answer: {answer}
History:
{history}
Please evaluate the confidence level of the current answer and provide a brief explanation.