M3CoTBench: Benchmark Chain-of-Thought of MLLMs in Medical Image Understanding

Anonymous authors
Paper under double-blind review

Abstract

Chain-of-Thought (CoT) reasoning has proven effective in enhancing large language models by encouraging step-by-step intermediate reasoning, and recent advances have extended this paradigm to Multimodal Large Language Models (MLLMs). In the medical domain, where diagnostic decisions depend on nuanced visual cues and sequential reasoning, CoT aligns naturally with clinical thinking processes. However, Current benchmarks for medical image understanding generally focus on the final answer while ignoring the reasoning path. An opaque process lacks reliable bases for judgment, making it difficult to assist doctors in diagnosis. To address this gap, we introduce a new M3CoTBench benchmark specifically designed to evaluate the correctness, efficiency, impact, and consistency of CoT reasoning in medical image understanding. M3CoTBench features ***1)*** a diverse, multi-level difficulty dataset covering **24** examination types, ***2)*** **13** varying-difficulty tasks, ***3)*** a suite of CoT-specific evaluation metrics (correctness, efficiency, impact, and consistency) tailored to clinical reasoning, and ***4)*** a performance analysis of multiple MLLMs. M3CoTBench systematically evaluates CoT reasoning across diverse medical imaging tasks, revealing current limitations of MLLMs in generating reliable and clinically interpretable reasoning, and aims to foster the development of transparent, trustworthy, and diagnostically accurate AI systems for healthcare.

1 Introduction

In recent years, Chain-of-Thought (CoT) reasoning has proven to be a transformative mechanism in enhancing the problem-solving capabilities of Large Language Models (LLMs) (Chu et al., 2024). By generating intermediate reasoning steps before arriving at a final answer, CoT improves transparency and structured decision-making in LLMs. Notable advancements include models like OpenAI's o1 (OpenAI, 2024b) and o3-mini (OpenAI, 2025), which exhibit consistent, step-by-step logical reasoning across multi-turn interactions, and DeepSeek-R1 (DeepSeek-AI et al., 2025) that excels at decomposing complex tasks into fine-grained subtasks. Building on these successes, researchers have extended CoT to Multimodal Large Language Models (MLLMs) (Wang et al., 2025), enabling joint processing of multiple modalities. Multimodal CoT (MCoT) frameworks now integrate visual and textual evidence into coherent multi-step explanations, with methods like Chain-of-Spot (Liu et al., 2024b), TextCoT (Luan et al., 2024), and DCoT (Jia et al., 2024) emphasizing region-of-interest analysis. Recent breakthroughs, such as OpenAI's o3 (OpenAI, 2024c) model, further demonstrate CoT's potential for image-based reasoning, while applications in healthcare, robotics, and autonomous driving highlight its versatility across domains.

In medical MLLMs, CoT reasoning is uniquely critical due to the complexity of medical image interpretation (Liu et al., 2024a). Clinicians rely on systematic diagnostic processes that involve iterative observation, verification against key features, and knowledge-based refinement. Explicit reasoning chains are essential to ensure safety, trustworthiness, and alignment with clinical guidelines. However, current medical imaging benchmarks focus solely on final-answer accuracy, neglecting the quality of intermediate reasoning steps (Wu et al., 2024; Ye et al., 2024; Hu et al., 2024). For instance, state-of-the-art Medical MLLM benchmarks evaluate VQA performance without assessing ***how*** or ***why*** a model arrives at an answer. This gap limits the development of clinically reliable AI systems, as two models could produce identical answers through fundamentally flawed or incomparable reasoning paths. Such a lack of scrutiny over intermediate reasoning increases the risk of unnoticed errors, misdiagnoses, and overconfidence in models that appear accurate on surface metrics.

To address these challenges, we introduce a novel M3CoTBench benchmark that is designed to evaluate and standardize CoT reasoning in medical image interpretation. Specifically, we propose a novel curation pipeline, which includes ***1)*** the collection of diverse and high-quality medical images, ***2)*** automated data annotation, and ***3)*** manual review and calibration. By bridging the gap between medical diagnostic workflows and AI-driven reasoning, M3CoTBench not only facilitates transparent evaluation but also paves the way for developing clinically trustworthy MLLMs. Our contributions redefine evaluation standards in medical imaging, emphasizing the need for interpretable, step-by-step reasoning in high-stakes applications. Our work is guided by three core principles:

- **Diverse Medical VQA Dataset.** We curate a 1,079-image QA dataset spanning 24 modalities, stratified by difficulty and annotated with step-by-step reasoning aligned to clinical workflows.
- **Multidimensional CoT-Centric Metrics.** Evaluation criteria for reasoning correctness, efficiency, impact, and consistency, enabling granular performance analysis for various MLLMs.
- **Comprehensive Model Analysis.** We evaluate general-purpose and medical MLLMs by quantitative metrics and case studies, highlighting strengths and failure modes in clinical reasoning to guide future improvements.

2 Related Work

2.1 Multimodal Large Language Models

Inspired by recent advances in large language models like LLaMA (Touvron et al., 2023) and GPT (Ouyang et al., 2022), MLLMs extend text-centric architectures by embedding visual features into the latent language space, enabling diverse image-grounded text generation. The LLaVA-OneVision (Li et al., 2024) family combines large-scale image/video corpora with instruction fine-tuning to excel across single-image, multi-image, and video tasks. LLaVA-CoT (Xu et al., 2024b) introduces a multistage prompting strategy incorporating summarization, visual analysis, reasoning, and conclusion. Qwen-2.5-VL (Bai et al., 2025) advances document parsing, diagram understanding, and step-by-step reasoning using dynamic resolution and temporal encoding. InternVL2.5 (Chen et al., 2024c) introduces a unified multimodal architecture with improved alignment and instruction-following capabilities across image and video inputs. Its tuned variant (Wang et al., 2024) further enhances CoT reasoning via multimodal preference optimization. Closed-source GPT-4o (OpenAI, 2024a) exemplifies integration of real-time vision, audio, and text reasoning. In medicine, specialized MLLMs adapt these techniques to clinical data: Med-Flamingo (Moor et al., 2023) augments Flamingo (Alayrac et al., 2022) with medical image–text pretraining for few-shot VQA; LLaVA-Med (Li et al., 2023) aligns visual content with biomedical concepts using PubMed captions and GPT-4 instructions; RadFM (Wu et al., 2023) pretrains on 2D/3D radiologic scans. Rapid progress demands more effective evaluation, underscoring the need for benchmarks targeting detailed diagnostic inference in complex multimodal contexts.

2.2 Medical Multimodal Benchmarks

Medical multimodal benchmarks evaluate how well MLLMs interpret and reason over clinical imaging data. VQA-RAD (Lau et al., 2018) is an early radiology VQA dataset with clinician-annotated QA pairs. PathVQA (He et al., 2020) extends VQA to pathology by pairing textbook and digital pathology images with expert-reviewed questions. SLAKE (Liu et al., 2021) offers English–Chinese radiology QA enriched with semantic labels linked to a structured medical knowledge base. FMBench (Wu et al., 2024) is the first to systematically assess fairness in MLLMs, incorporating clinical tasks, demographic-aware evaluation, and a novel disparity metric. Quilt-VQA (Seyfioglu et al., 2024) targets histopathology VQA using real-world images and curated questions. OmniMedVQA (Hu et al., 2024) aggregates diverse datasets spanning multiple modalities and anatomy, requiring models to integrate heterogeneous inputs and justify their answers. GMAI-MMBench (Ye et al., 2024) unifies 284 global datasets into a large-scale multimodal QA benchmark covering a broad range of clinical scenarios. Despite these advances, most benchmarks still focus on surface-level Q&A and rarely evaluate deep diagnostic reasoning, such as inferring disease etiology or treatment decisions from imaging findings. Moreover, they often lack annotations for intermediate reasoning steps, limiting their effectiveness in assessing CoT-style clinical inference.

2.3 CoT-Related MLLM Benchmarks

Research on reasoning in multimodal models has advanced through several dedicated benchmarks. Visual-CoT (Shao et al., 2024) introduces a large-scale dataset of image–Q&A pairs, augmented with region annotations and step-by-step rationales, along with a multi-turn reasoning pipeline for interpretable, region-focused CoT tasks. M^{3}CoT (Chen et al., 2024b) provides a comprehensive benchmark spanning diverse domains and requiring complex multi-step visual–textual reasoning. MME-CoT (Jiang et al., 2025) extends this line of work by contributing high-quality data across six domains and proposing three novel metrics to assess CoT quality, robustness, and efficiency. CoMT (Cheng et al., 2025) introduces a benchmark that requires both multimodal inputs and outputs to evaluate the visual reasoning abilities of LVLMs, addressing the limitations of traditional text-only outputs in multimodal CoT tasks. MMIR (Yan et al., 2025) is designed to evaluate MLLMs' ability to detect and reason about semantic inconsistencies in layout-rich multimodal content, revealing significant shortcomings in current models and highlighting the need for more advanced cross-modal reasoning capabilities. While these benchmarks have advanced CoT reasoning in natural image domains, analogous resources remain scarce in the medical field, where rigorous diagnostic reasoning, interpretability, and domain expertise are essential. This gap underscores the need for medically grounded benchmarks that can assess step-by-step clinical inference in multimodal settings.

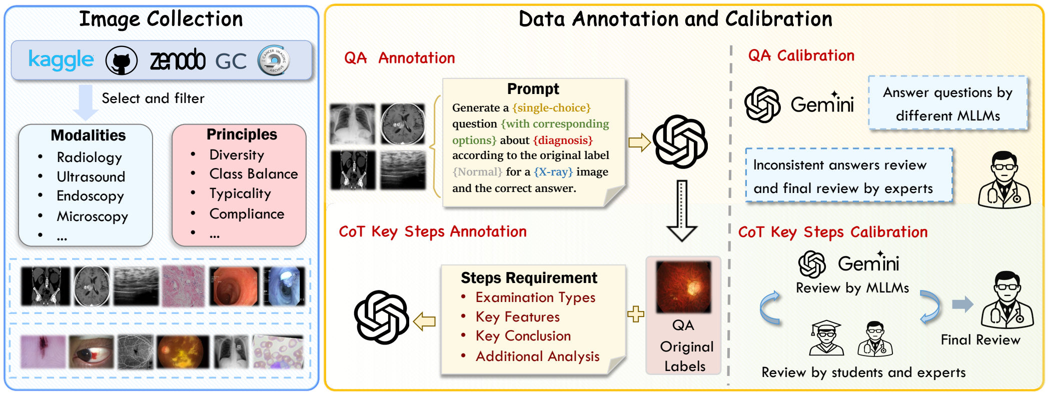


Figure 1: **Curation of M3CoTBench benchmark** that encompasses three sections: ***1)*** carefully curated medical images from various public sources, ***2)*** multi-type and multi-difficulty QA generation via LLMs and expert calibration, ***3)*** and structured annotation of key reasoning steps aligned with clinical diagnostic workflows.

3 Curation of M3CoTBench

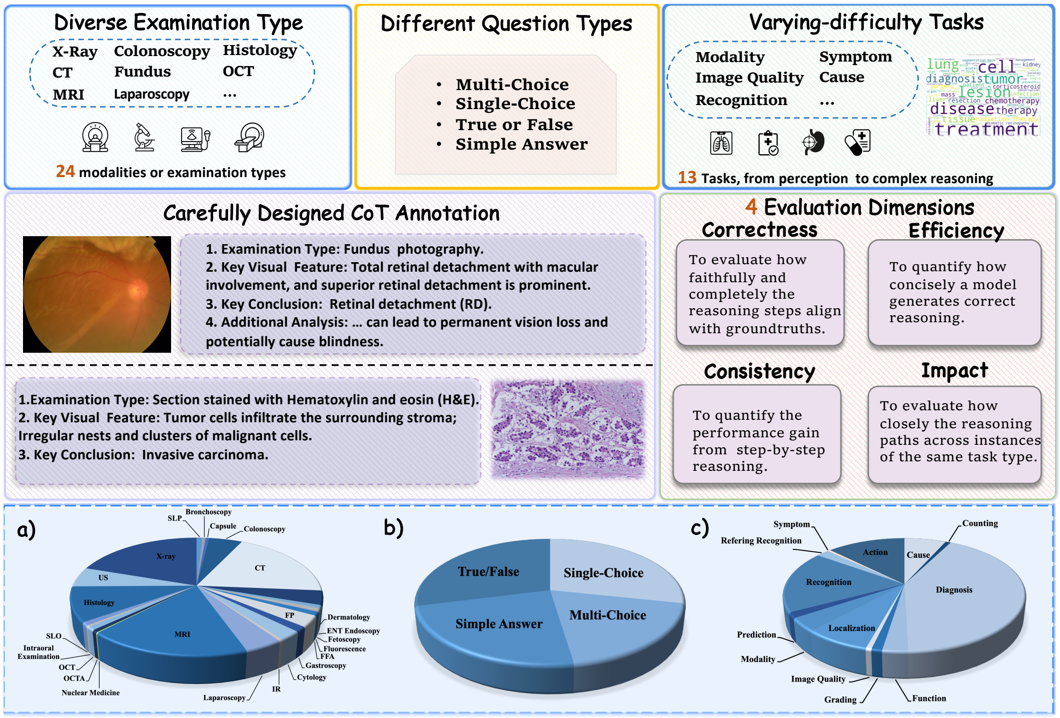
The collection of images, construction of QA pairs, the annotation of key CoT steps, and manual review/calibration are carefully designed in Figure 1.

3.1 Data Collection

All images in M3CoTBench are sourced from public datasets, with selection guided by principles of diversity, representativeness, class balance, and compliance.

- **Diversity.** Images are collected from 55 public medical datasets, encompassing diverse imaging modalities, examination types, and anatomical regions (Table A1), with broad geographical coverage (Figure A3) and diverse temporal ranges of publishing.
- **Typicality.**To ensure large intra-dataset variance, image features are extracted by Biomed-CLIP (Zhang et al., 2023), and a semantically distinct subset is selected by maximizing the minimum pairwise feature distance.
- **Class balance.** Each dataset includes multiple categories, with a balanced class distribution maintained through manual review based on original labels.

- **Compliance.** Datasets with usage restrictions or labeled as “no derivatives” are excluded, addressing compliance issues often neglected in prior benchmarks.



1. Examination Type: Fundus photography.
2. Key Visual Feature: Total retinal detachment with macular involvement, and superior retinal detachment is prominent.
3. Key Conclusion: Retinal detachment (RD).
4. Additional Analysis: … can lead to permanent vision loss and potentially cause blindness.

1.Examination Type: Section stained with Hematoxylin and eosin (H&E).
2. Key Visual Feature: Tumor cells infiltrate the surrounding stroma; Irregular nests and clusters of malignant cells.
3. Key Conclusion: Invasive carcinoma.

Correctness: To evaluate how faithfully and completely the reasoning steps align with groundtruths.

Efficiency: To quantify how concisely a model generates correct reasoning.

Consistency: To quantify the performance gain from step-by-step reasoning.

Impact: To evaluate how closely the reasoning paths across instances of the same task type.

Figure 2: **Overview of M3CoTBench benchmark. Top:** The benchmark covers 24 imaging modalities/examination types, 4 question types, and 13 clinical reasoning tasks. **Middle:** CoT annotation examples and 4 evaluation dimensions. **Bottom:** The distribution of image-QA pairs across *a)* modalities, *b)* question types, and *c)* tasks.

3.2 Data Annotation and Calibration

Question-Answer Pairs Generation. We employ a unified pipeline for generating QA pairs, with all questions and candidate answers initially fully generated by GPT-4o, and subsequently calibrated by three different MLLMs and human experts to ensure the validity of the questions and the correctness of the answers.

- **Conversion of Existing Datasets.** To accommodate the original purpose of each dataset, we tailor strategies to different data types. Starting with existing QA pairs from public VQA and image classification datasets, we use GPT-4o to rewrite them into more diverse formats, such as single-choice, multiple-choice, true/false, and short-answer questions. For segmentation datasets, we concatenate the raw image with its corresponding mask and ask targeted questions about the masked region; for object detection datasets, we generate spatial questions, such as requesting a rough anatomical location or estimating bounding box coordinates; and for image quality assessment and disease grading tasks, we present paired images and formulate comparative questions.
- **Generation of Inference-driven Medical Questions.** To enrich the complexity of QA tasks and better support reasoning capabilities, we provide GPT-4o with the original label and prompt it to generate questions with corresponding answer options grounded in that information. For example, given a slit lamp image labeled ”severe keratitis with corneal ulcer”, GPT-4o is prompted to create a multi-choice question about causes, such as “What might be the cause of this condition? (Select all that apply)”, with answer options including bacterial, viral, fungal infections, trauma, allergic reactions, etc. The correct answers align with clinically relevant causes associated with the diagnosis. This approach introduces hierarchical difficulty and inference-driven tasks that go beyond surface-level recognition, fostering deeper medical reasoning.

- **AI and Human Expert Calibration Process.** For calibration, we leverage three different MLLMs to answer each image-question pair independently. If any MLLM's response differs from the initially generated answer, a human expert, an experienced doctor, intervenes to make the final judgment. Additionally, the expert reviews all images and QA pairs comprehensively to perform a final quality check and calibration. This combined AI-human validation ensures high accuracy and reliability of the dataset.

Rationale for the step design. Our CoT steps, (1) confirming the image's nature (modality/examination type), (2) identifying key visual features, (3) drawing diagnostic conclusions, and (4) providing further medically informed analysis, are derived from clinician interviews and established theories of medical reasoning.

- **Validation via doctor interviews.** Before designing the CoT steps, we interviewed clinicians, radiologists, and sonographers from five hospitals. Most described their workflow as: identify the imaging modality, observe key features, draw core conclusions, and then perform additional analyses such as etiology or treatment planning. One doctor noted that intuition may guide an initial hypothesis, which is then verified through feature inspection. These findings support our chosen steps as both sufficient and necessary for medical reasoning.
- **Theoretical support from medical cognition.** Our CoT design draws on established cognitive models. a) Hypothetico-deductive reasoning (Elstein et al., 1978): Clinicians generate and iteratively test hypotheses; our steps follow this natural cycle. b) Pattern recognition (Norman et al., 2007): Experienced doctors rapidly spot salient imaging patterns; our early focus on key features reflects this process. c) Dual-process theory (Arvai, 2013): Intuitive and analytical reasoning interact; our annotations capture this by allowing preliminary intuitive judgments followed by feature-based verification and further analysis.

CoT key steps Generation. To ensure effective CoT in medical VQA that mirrors clinicians' cognitive workflow from perception to judgment, we first leverage MLLMs to annotate CoT key steps, which then undergo repeated cycles of review, feedback, and revision by medical experts and students before senior experts confirm the final CoT annotations.

- **MLLM-Based Annotation.** For each image-question-answer instance, we provide GPT-4o and Gemini-2.5-Pro with the image, the question, the answer, and any relevant contextual information from the original annotations. For example, underlying labels used to construct the question itself, complex questions about treatment, causes, prediction, or function are often derived from simpler labels such as disease type, which are also provided as input. Additionally, the model generates reasoning steps following an expert-designed four-step clinical structure: (1) confirming the nature of the image, such as the imaging modality and examination type; (2) identifying key visual features; (3) drawing diagnostic conclusions, including the relevant disease, organ, or tissue; and (4) providing additional analysis based on medical knowledge, such as treatment strategies or associated symptoms. It is worth noting that we condition the model by specifying the expected reasoning steps based on the task type. For instance, modality questions omit steps (3) and (4), while diagnostic questions skip step (4). GPT-4o and Gemini-2.5-Pro then generate the corresponding key reasoning steps accordingly. Finally, the final results were generated again by GPT-4o, which integrated annotation information from both GPT-4o and Gemini-2.5-Pro.
- **AI and Human Expert Calibration Process.** To ensure high-quality and medically reliable annotations, we adopt a multi-stage human–AI collaborative verification process: a) Initial Student Review: A medically trained student manually reviews model- or human-generated annotations, correcting factual, spelling, and formatting errors, and filling in missing key information. Uncertain cases are discussed with experts. b) Automated Multi-Model Checking: The image, question, and reasoning steps are validated using GPT-4o. c) Expert Review on Model Flags: Any reasoning step flagged as "potentially incorrect" by any model is sent to an expert in the relevant imaging modality for manual review. d) Consensus Resolution: When experts identify issues, the involved experts and student reviewers hold brief online or asynchronous discussions to resolve disagreements. Three such meetings and multiple asynchronous discussions were held. Final reasoning steps and conclusions are updated based on consensus. e) Final Expert Read-through: Experts conduct a final pass on each sample to ensure that the image, question, reasoning chain, and answer are medically correct, consistent, and compliant with benchmark standards.

3.3 Data Composition and Categorization

As shown in Figure 2, M3CoTBench includes diverse image–QA pairs with multiple question formats and task types of varying difficulty. It covers a broad range of imaging modalities across several categories. Tasks span from basic perception to advanced medical reasoning, enabling comprehensive evaluation of MLLMs.

QA Types. We include four question formats: single-choice, multiple-choice, true/false (judgment), and short-answer, spanning 13 task types with varying difficulty levels.

Examination Types. The dataset encompasses 24 imaging modalities and examination methods, which can be organized into six major categories: ophthalmic imaging, radiology, endoscopy, microscopy, ultrasound-based examinations, and surface-level inspections. Representative modalities within these categories include slit lamp photography (SLP), fundus photography (FP), optical coherence tomography (OCT), optical coherence tomography angiography (OCTA), scanning laser ophthalmoscopy (SLO), fundus fluorescein angiography (FFA), X-ray, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound (US), infrared reflectance (IR), nuclear medicine, fetoscopy, laparoscopy, colonoscopy, gastroscopy, capsule endoscopy, bronchoscopy, ENT endoscopy, cytology, fluorescence microscopy, dermoscopy, and intraoral examination.

Task Types. To thoroughly assess the reasoning ability of MLLMs, we design questions spanning a broad spectrum of clinical tasks, including: Examination Type, Image Quality, Recognition, Referring Recognition, Localization, Diagnosis, Grading, Prediction, Function, Symptom, Counting, Cause, and Action. These categories range from low-level perception tasks (e.g. assessing image quality) to high-level clinical reasoning (e.g. identifying causal factors or suggesting next actions). Such a taxonomy is constructed to test MLLMs' ability to bridge the gap between visual perception and domain knowledge reasoning, challenging both their vision-language alignment and medical understanding. Some example image-question pairs can be seen in Figure 3.

Table 1: **Criterion comparison for current benchmarks. ✓**: Satisfied. **✗**: Unsatisfied.

Dataset	**#Img/#QA**	**Exam. Type**	**Task**	**Question Type**	**CoT Annotation**	**Eval. Dimension**			
						Corr.	Imp.	Eff.	Cons.
VQA-RAD (Lau et al., 2018)	315 / 3515	3	8	2	✗	✓	✗	✗	✗
SLAKE (Liu et al., 2021)	642 / 14028	3	10	2	✗	✓	✗	✗	✗
Quilt-VQA (Seyfioglu et al., 2024)	985 / 1283	2	5	2	✗	✓	✗	✗	✗
OmniMedVQA (Hu et al., 2024)	118010 / 127995	12†	5	1	✗	✓	✗	✗	✗
GMAI-MMBench (Ye et al., 2024)	- / 25831	**38**†	6	1	✗	✓	✗	✗	✗
M3CoTBench	1079 / 1079	24	**13**	**4**	✓	✓	✓	✓	✓

† The way of classifying modalities differs from this paper.

4 Evaluation Suite of M3CoTBench

We evaluate CoT reasoning based on four aspects: correctness, efficiency, impact, and consistency. Here, correctness measures whether the generated reasoning steps are accurate; efficiency reflects the additional inference time introduced by reasoning; impact quantifies the overall effect of reasoning on answer accuracy compared to direct prediction without reasoning; and consistency assesses whether similar tasks tend to follow similar reasoning paths.

Evaluation of Reasoning Correctness. To comprehensively evaluate the accuracy of the model's reasoning steps, we quantify the alignment between the generated reasoning sequence and expert-annotated reasoning paths. Specifically, we compute the following metrics:

$$\text{Avg Precision} = 1/N \sum_{i=1}^{N} |\mathcal{R}^{(i)} \cap \mathcal{A}_{k^*}^{(i)}| / |\mathcal{R}^{(i)}|, \ \text{Avg Recall} = 1/N \sum_{i=1}^{N} |\mathcal{R}^{(i)} \cap \mathcal{A}_{k^*}^{(i)}| / |\mathcal{A}_{k^*}^{(i)}|. \quad (1)$$

Here, $\mathcal{R}$ denotes the set of reasoning steps generated by the model, and $\{\mathcal{A}_k\}$ represents all annotated gold reasoning paths for a given question. Since multiple valid reference paths may exist, we choose the reference $\mathcal{A}_{k^*}$ with the highest overlap with $\mathcal{R}$. Precision measures the proportion of model-generated steps that are correct, while recall quantifies the coverage of reference reasoning steps. The F1 score is used to combine both aspects to provide a holistic evaluation of CoT correctness.

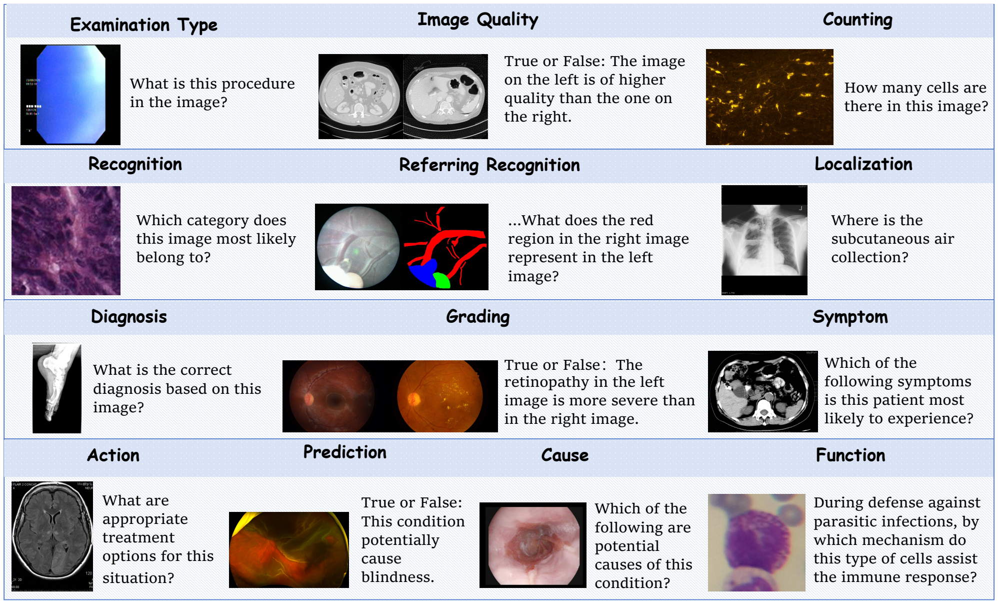

Examination Type
What is this procedure in the image?
Image Quality
True or False: The image on the left is of higher quality than the one on the right.
Counting
How many cells are there in this image?
Recognition
Which category does this image most likely belong to?
Referring Recognition
...What does the red region in the right image represent in the left image?
Localization
Where is the subcutaneous air collection?
Diagnosis
What is the correct diagnosis based on this image?
Grading
True or False: The retinopathy in the left image is more severe than in the right image.
Symptom
Which of the following symptoms is this patient most likely to experience?
Action
What are appropriate treatment options for this situation?
Prediction
True or False: This condition potentially cause blindness.
Cause
Which of the following are potential causes of this condition?
Function
During defense against parasitic infections, by which mechanism do this type of cells assist the immune response?

Figure 3: **Example image-question pairs for 13 tasks in M3CoTBench**, including identifying examination types, image quality assessment, recognition, referring recognition, counting, localization, diagnosis, grading, symptom identification, clinical action planning, prediction, functional understanding, and causal reasoning

Evaluation of Reasoning Efficiency. CoT reasoning often introduces significant computational overhead due to longer generated sequences. Excessively verbose CoT outputs increase inference time and memory consumption, reducing practical usability in real-world applications. To evaluate reasoning efficiency, we compute the number of correct reasoning steps per unit time. Formally,

$$E = (|\mathcal{R} \cap \mathcal{A}_{k^*}|)/T_{\text{CoT}}. \tag{2}$$

A higher E indicates more accurate reasoning steps per unit time, reflecting more efficient reasoning. Then we define the average inference latency impact L as the difference between the total CoT inference time and total direct inference time, divided by the number of examples: $L = T_{\text{CoT}}/T_{\text{direct}}$, where N is the number of examples, and T_{CoT}, T_{direct} are the total inference times with and without CoT, respectively. A larger L value indicates a greater average latency overhead. By jointly considering E and L, we can better benchmark the trade-offs between interpretability and computational cost in CoT-enabled models.

Evaluation of Reasoning Impact. To quantify the benefit of generating step-by-step reasoning over directly producing the final answer, we define the reasoning impact metric as the difference in answer accuracy between the two approaches. Let Acc_{step} denote the accuracy of the model when generating answers with intermediate reasoning steps, and $\text{Acc}_{\text{direct}}$ denote the accuracy when generating answers directly without explicit reasoning. The reasoning impact I is computed as: $I = \text{Acc}_{\text{step}} - \text{Acc}_{\text{direct}}$.

A positive value of I indicates that step-by-step reasoning improves answer correctness, demonstrating the effectiveness of CoT generation in enhancing model performance. Conversely, a negative or zero value suggests that the reasoning steps do not provide additional benefit or may even degrade the correctness. This metric offers a straightforward way to assess whether incorporating explicit reasoning contributes meaningfully to the model's accuracy.

Evaluation of Reasoning Consistency. Structured and task-specific reasoning pathways are fundamental for interpretability and play a vital role in ensuring reproducibility, transparency, and trustworthiness in high-stakes medical decision-making. However, existing evaluation metrics often treat reasoning steps as unordered elements. To address this gap, we introduce a path consistency metric that explicitly evaluates how closely the reasoning path for each instance in the same task type. We compute this score independently for each of the thirteen tasks and then average the results. For task t with N examples, represent each generated reasoning path $P_i^{(t)}$ as an ordered sequence of step categories (e.g. *modality*,*feature*, *diagnosis*, *additional analysis*).To evaluate path consistency, we first select the reference path by maximizing its average similarity with all generated paths:

$$P^{(t)} = \arg\max_{P} \sum_{i=1}^{N} \mathrm{sim}\big(P, P_i^{(t)}\big), \tag{3}$$

$$\mathrm{sim}\big(P, P_i^{(t)}\big) = |\mathrm{LCS}(P, P_i^{(t)})|/\max(|P|, |P_i^{(t)}|). \tag{4}$$

The task-level consistency score is then defined as the average similarity between each path and the canonical reference:

$$C_{\mathrm{path}}^{(t)} = 1/N \sum_{i=1}^{N} \mathrm{sim}(P^{(t)}, P_i^{(t)}). \tag{5}$$

Average score over all tasks: $C_{\mathrm{path}} = 1/M \sum_{t=1}^{M} C_{\mathrm{path}}^{(t)}$, where $M = 13$. A higher $C_{\mathrm{path}} \in [0, 1]$ indicates that the model has strong structural stability in its CoT.

5 Experiments

5.1 Experiment Setup

Evaluation Models. We select top-performing MLLMs for comprehensive CoT evaluation. We test models such as LLaVA-OneVision(7B) (Li et al., 2024), Qwen2.5-VL (7B,72B) (Bai et al., 2025), Llama-3.2-Vision-Instruct(11B, 90B) (Meta AI, 2024), which are not trained for the reasoning capability. We also include closed-source GPT-4o (OpenAI, 2024a) and Gemini 2.5 Pro (Google DeepMind, 2024) as a strong baseline model. Besides, we test recent models targeting reasoning like LLaVA-CoT (11B) (Xu et al., 2024a). Finally, we evaluate some models specifically designed for the medical domain, like LLaVA-Med (7B) (Li et al., 2023), HuatuoGPT-Vision-7B-Qwen2.5VL (Chen et al., 2024a) and HealthGPT (Lin et al., 2025).

Implementation Details. We define the CoT prompt as: *Please generate a step-by-step answer, including all intermediate reasoning steps, and provide the final answer at the end.* The direct prompt is defined as: *Please directly provide the final answer without any additional output.* For all experiments, the batch size is set to 1 to ensure independent processing of each sample, and the temperature is uniformly set to 0.1. For evaluation, we use GPT-4o for all assessment criteria. All local inference experiments were conducted on a server with NVIDIA H20 GPUs. APIs are used for closed-source MLLMs, Qwen2.5-VL-Instruct, and Llama-3.2-Vision series.

5.2 Quantitative Results

The experimental results can be seen in Table 2, from which there are some interesting findings:

Correctness. LLaVA-CoT exhibits relatively strong performance under the CoT setting, likely due to its architecture and training process, which emphasize structured reasoning chains while minimizing irrelevant or misleading steps. This design helps preserve accuracy and suggests that CoT effectiveness depends not only on prompt structure but also on a model's inherent ability to generate reliable intermediate reasoning. In contrast, medical-specific models such as LLaVA-Med and HuatuoGPT-Vision show much lower correctness scores, indicating limitations in generalizing to complex reasoning tasks beyond domain-specific patterns. Importantly, the correctness of CoT reasoning is closely tied to overall performance: aside from LLaVA-Med, which performs poorly even without CoT, models producing high-accuracy CoT tend to suffer less degradation when CoT is applied. This implies that effective CoT designs can improve medical image understanding and reasoning by enhancing the model's ability to structure and verify intermediate steps.

Efficiency. After introducing step-by-step reasoning prompts, models show markedly different latency behaviors. LLaVA-CoT, the only open-source model explicitly optimized for CoT reasoning, experiences minimal additional delay. Some closed-source models show moderate, acceptable latency increases due to a few extra decoding steps. In contrast, Qwen2.5-VL-7B-Instruct, experiences a substantial increase in latency, likely due to repeated processing of visual inputs and the lack of

Table 2: **M3CoTBench results for MLLMs.** ↑(↓): the higher(lower) the better. $F1, P, R$: the average of F1 score(%), Precision(%), and Recall(%). $\text{Acc}_{\text{direct}}$ and Acc_{step}: accuracy(%) of generated answers by directly and CoT. I, E, L, and $C_{\mathtt{path}}$: Impact, Efficiency, Latency, and Consistency score, respectively. Optimal / sub-optimal results are highlighted in **bold** / underline.

Model	Correctness			Impact			Efficiency		Consistency
	$F1(\uparrow)$	$P(\uparrow)$	$R(\uparrow)$	Acc_{step}	$\text{Acc}_{\text{direct}}$	$I(\uparrow)$	$E(\uparrow)$	$L(\downarrow)$	$C_{\mathtt{path}}(\uparrow)$
Open-source MLLMs									
LLaVA-OV-7B (Li et al., 2024)	39.95	39.33	40.60	34.85	41.80	-6.95	16.36	11.06	0.783
LLaVA-CoT (11B) (Xu et al., 2024a)	**61.27**	**71.36**	53.68	40.59	40.69	**-0.10**	22.38	1.35	0.630
Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	41.92	37.06	48.26	35.13	43.93	-8.80	18.85	15.45	0.822
Qwen2.5-VL-72B-Instruct (Bai et al., 2025)	50.03	44.58	57.01	46.25	55.24	-8.99	13.55	7.76	0.853
Llama-3.2-11B-Vision (Meta AI, 2024)	36.97	32.34	43.14	39.85	44.21	-4.36	13.97	1.13†	0.823
Llama-3.2-90B-Vision (Meta AI, 2024)	47.72	39.26	47.72	42.63	51.81	-9.18	11.95	5.30	0.811
Closed-source MLLMs									
Gemini 2.5 Pro (Google DeepMind, 2024)	57.08	46.53	**73.82**	**57.10**	**58.60**	-1.50	9.95	1.68	0.835
Claude-Sonnet-4 (Anthropic, 2024)	56.31	52.09	61.28	45.06	46.34	-1.28	17.01	2.68	**0.871**
GPT-4o (OpenAI, 2024a)	54.46	50.70	58.84	49.85	52.64	-2.79	13.32	5.81	0.834
GPT-4.1 (OpenAI, 2023)	54.41	45.26	68.21	54.82	55.82	-1.00	19.53	4.88	0.863
Medical MLLMs									
LLaVA-Med(7B) (Li et al., 2023)	31.85	37.98	27.43	27.99	28.36	-0.37	**37.68**	2.67	0.660
HuatuoGPT-Vision-7B (Chen et al., 2024a)	33.70	32.34	35.19	33.64	43.47	-9.83	15.21	20.97	0.833
HealthGPT(3.8B) (Lin et al., 2025)	53.79	53.21	54.38	41.24	43.93	-2.69	16.82	8.33	0.578

† Due to issues related to the API, the inference speed of Llama-3.2-11B-Vision is particularly slow.

embedding caching. HuatuoGPT-Vision-7B-Qwen2.5 exhibits the largest slowdown, likely because each reasoning step redundantly triggers the full vision-language pipeline. Overall, explicit CoT support and visual embedding reuse emerge as key factors for efficient CoT execution.

Impact. In this accuracy comparison, closed-source models generally outperform their open-source counterparts, with Gemini 2.5 Pro and GPT-4.1 achieving the highest scores. Among open-source systems, Qwen2.5-VL-72B-Instruct stands out, delivering performance close to proprietary models. Most medical-specific models lag behind, reflecting limited generalization in complex medical reasoning tasks. Notably, CoT prompting fails to yield consistent gains in medical image understanding and can even slightly reduce precision, likely because it introduces unnecessary or misleading reasoning steps in domains where diagnostic decisions depend more on visual cues than logical inference. The problem is especially pronounced when medical models lack robust multimodal grounding, and CoT may further raise hallucination risk or distract attention from critical features, as recently noted in (Li et al., 2025). Some prior studies have discussed this phenomenon. (Mishra & Thakkar, 2023) points out that CoT is highly sensitive, and unreasonable reasoning chains may substantially degrade performance. The effects of CoT in (Jiang et al., 2025) are measured: most perception tasks showed decreased performance, while about half of the reasoning tasks declined. In some open-ended medical VQA tasks, enabling CoT in Gemini-2.5-Flash resulted in worse performance than non-CoT mode, with a drop of 1.28% (Hong et al., 2025). Interestingly, LLaVA-CoT shows the smallest accuracy drop with CoT, likely due to its reasoning design that avoids irrelevant steps, while closed-source models remain stable thanks to stronger multimodal fusion and richer training data. Making CoT genuinely effective for medical image understanding remains an open challenge.

Consistency. Apart from models like LLaVA-CoT, LLaVA-Med, and HealthGPT, most models show relatively high consistency, following similar reasoning paths on the same tasks. Closed-source models are particularly consistent, likely due to larger, carefully curated datasets, rigorous fine-tuning, and strict output and reasoning protocols. These factors help ensure their outputs are accurate, stable, and repeatable, reducing variability in intermediate steps and final answers, which is especially important in high-stakes medical tasks.

5.3 Qualitative Analysis

By analysis of model outputs with errors, systematic errors are emerging within the intermediate steps in CoT rather than merely at the final prediction. For example, in the pathology question the CoT output misclassified the case as *Dysplasia* even though the early reasoning correctly noted "cellular atypia", but then failed to verify the key criterion of "full-thickness epithelial involvement with an intact basement membrane". Such qualitative inspection highlights three factors:

1. **Incomplete verification of decisive diagnostic features.** Although the CoT reasoning often identified some relevant abnormalities, it frequently omitted or misweighted critical criteria, such as the extent of epithelial involvement in the pathology case, thereby allowing early misreadings to dominate the conclusion and persist through the subsequent steps.
2. **Weakened vision-language grounding during step-wise verbalization.** By forcing the model to translate visual cues into descriptive textual representations before decision-making, CoT increased the risk of information distortion, subtle semantic drift, and gradual loss of fine visual detail. In the hematology example, this intermediate translation process led to an inaccurate verbal focus on nuclear shape while neglecting the defining cytoplasmic granules, their relative prominence, and characteristic spatial distribution.
3. **Error accumulation along the reasoning chain.** Once an early descriptive mistake occurred, subsequent steps propagated and rationalized the error, producing a seemingly coherent but ultimately incorrect explanation that became harder to override with additional context.

These observations indicate that the degradation under CoT reflects deeper vulnerabilities in how visual evidence is interpreted and verified across multiple reasoning stages. Representative examples and detailed error analyses are provided in Appendix D.2.

6 Conclusion

In this work, we introduce M3CoTBench, a novel benchmark designed to evaluate CoT reasoning in MLLMs for medicine. Our benchmark addresses the critical gap between answer correctness and reasoning quality in clinical AI systems by incorporating diverse imaging modalities or examination types, step-by-step reasoning annotations, and tailored multi-dimensional evaluation metrics across medical cases of varying difficulty, from simple pattern recognition to complex diagnostic reasoning, enabling fine-grained analysis of model capabilities. Through comprehensive assessments of state-of-the-art MLLMs, we demonstrate limitations of existing models in generating interpretable and clinically aligned reasoning. We hope this benchmark will inspire future research toward more transparent, trustworthy, and practically valuable AI systems for healthcare and beyond. More discussions about limitations and social impact can be seen in Appendix E and Appendix F.

Ethics statement

We have ensured that our study and dataset construction follow ethical standards, with no direct involvement of human subjects, and no foreseeable risk of harm. Data usage complies with privacy and legal requirements, and we have aimed to mitigate potential biases in annotations and model evaluation. We disclose no conflicts of interest or sponsorship that could influence the results.

Reproducibility Statement

We have already elaborated on all the models or algorithms proposed, experimental configurations, and benchmarks used in the experiments in the main body or appendix of this paper. Furthermore, we declare that the entire code used in this work will be released after acceptance.

References

Covid-19 lung ct scans. https://www.kaggle.com/datasets/luisblanche/covidct/data, 2020.

Entrep challenge: Advancing vision-language ai for ent endoscopy analysis, 2025. URL https://aichallenge.hcmus.edu.vn/acm-mm-2025/entrep.

Maruf Adewole, Jeffrey D. Rudie, Anu Gbadamosi, Oluyemisi Toyobo, Confidence Raymond, Dong Zhang, Olubukola Omidiji, Rachel Akinola, Mohammad Abba Suwaid, Adaobi Emegoakor, Nancy Ojo, Kenneth Aguh, Chinasa Kalaiwo, Gabriel Babatunde, Afolabi Ogunleye, Yewande Gbadamosi, Kator Iorpagher, Evan Calabrese, Mariam Aboian, Marius Linguraru, Jake Albrecht, Benedikt Wiestler, Florian Kofler, Anastasia Janas, Dominic LaBella, Anahita Fathi Kzerooni, Hongwei Bran Li, Juan Eugenio Iglesias, Keyvan Farahani, James Eddy, Timothy Bergquist, Verena Chung, Russell Takeshi Shinohara, Walter Wiggins, Zachary Reitman, Chunhao Wang, Xinyang Liu, Zhifan Jiang, Ariana Familiar, Koen Van Leemput, Christina Bukas, Maire Piraud, Gian-Marco Conte, Elaine Johansson, Zeke Meier, Bjoern H Menze, Ujjwal Baid, Spyridon Bakas, Farouk Dako, Abiodun Fatade, and Udunna C Anazodo. The brain tumor segmentation (brats) challenge 2023: Glioma segmentation in sub-saharan africa patient population (brats-africa), 2023.

Moulay A. Akhloufi and Mohamed Chetoui. Chest XR COVID-19 detection. https://cxr-covid19.grand-challenge.org/, August 2021. Online; accessed September 2021.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Shams Nafisa Ali, Md. Tazuddin Ahmed, Tasnim Jahan, Joydip Paul, S. M. Sakeef Sani, Nawshaba Noor, Anzirun Nahar Asma, and Taufiq Hasan. A web-based mpox skin lesion detection system using state-of-the-art deep learning models considering racial diversity. *Biomedical Signal Processing and Control*, 98:106742, 2024.

Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019.

Anthropic. Claude sonnet 4. https://www.anthropic.com/index/claude, 2024. Large Language Model.

Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, et al. Bach: Grand challenge on breast cancer histology images. *Medical Image Analysis*, 56:122–139, August 2019. ISSN 1361-8415. doi: 10.1016/j.media.2019.05.010. URL http://dx.doi.org/10.1016/j.media.2019.05.010.

Joseph Arvai. Thinking, fast and slow, daniel kahneman, farrar, straus & giroux, 2013.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Sophia Bano, Francisco Vasconcelos, Luke M Shepherd, Emmanuel Vander Poorten, Tom Vercauteren, Sebastien Ourselin, Anna L David, Jan Deprest, and Danail Stoyanov. Deep placental vessel segmentation for fetoscopic mosaicking. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pp. 763–773. Springer, 2020.

Sophia Bano, Alessandro Casella, Francisco Vasconcelos, Sara Moccia, George Attilakos, Ruwan Wimalasundera, Anna L David, Dario Paladini, Jan Deprest, Elena De Momi, et al. Fetreg: Placental vessel segmentation and registration in fetoscopy challenge dataset. *arXiv preprint arXiv:2106.05923*, 2021.

Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Working Notes of CLEF 2019*, volume 2380 of *CEUR Workshop Proceedings*, Lugano, Switzerland, September 9-12 2019. CEUR-WS.org. URL https://ceur-ws.org/Vol-2380/paper_272.pdf.

Sartaj Bhuvaji, Ankita Kadam, Prajakta Bhumkar, and Sameer Dedge. Brain tumor classification (mri). https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri, 2020.

Federico Bolelli, Luca Lumetti, Shankeeth Vinayahalingam, Mattia Di Bartolomeo, Arrigo Pellacani, et al. Segmenting the Inferior Alveolar Canal in CBCTs Volumes: the ToothFairy Challenge. *IEEE Transactions on Medical Imaging*, pp. 1–17, Dec 2024. ISSN 1558-254X. doi: https://doi.org/10.1109/TMI.2024.3523096.

Mateusz Buda, Ashirbani Saha, and Maciej A Mazurowski. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in biology and medicine*, 109:218–225, 2019.

Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale, 2024a. URL https://arxiv.org/abs/2406.19280.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8199–8221, 2024b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024c.

Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23678–23686, 2025.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *ACL*, 2024.

Marco Cipriano, Stefano Allegretti, Federico Bolelli, Federico Pollastri, and Costantino Grana. Improving Segmentation of the Inferior Alveolar Nerve through Deep Label Propagation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21105–21114. IEEE, Jun 2022. ISBN 978-1-6654-6947-0. doi: https://doi.org/10.1109/CVPR52688.2022.02046.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li,

Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Jianning Deng, Peize Li, Kevin Dhaliwal, Chris Xiaoxuan Lu, and Mohsen Khadem. Feature-based visual odometry for bronchoscopy: A dataset and benchmark. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6557–6564. IEEE, 2023.

Simon Duchesne, Isabelle Chouinard, Olivier Potvin, Vladimir S Fonov, April Khademi, Robert Bartha, Pierre Bellec, D Louis Collins, Maxime Descoteaux, Rick Hoge, et al. The canadian dementia imaging protocol: harmonizing national cohorts. *Journal of Magnetic Resonance Imaging*, 49(2):456–465, 2019.

Arthur S Elstein, Lee S Shulman, and Sarah A Sprafka. *Medical problem solving: An analysis of clinical reasoning*. Harvard University Press, 1978.

Mohammad Fraiwan, Ziad Audat, Luay Fraiwan, and Tarek Manasreh. Using deep transfer learning to detect scoliosis and spondylolisthesis from x-ray images. *Plos one*, 17(5):e0267851, 2022.

Wilfrido Gómez-Flores, Maria Julia Gregorio-Calas, and Wagner Coelho de Albuquerque Pereira. Bus-bra: a breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical Physics*, 51(4):3110–3123, 2024.

Google DeepMind. Gemini 2.5 pro. https://deepmind.google/technologies/gemini/, 2024. Accessed: May 2025.

Shivanand Gornale and Pooja Patravali. Digital knee x-ray images. *Mendeley Data*, 1, 2020.

Hayden Gunraj, Chi en Amy Tai, and Alexander Wong. Cancer-net pca-data: An open-source benchmark dataset for prostate cancer clinical decision support using synthetic correlated diffusion imaging data. *NeurIPS Workshops*, 2023.

Palak Handa, Amirreza Mahbod, Florian Schwarzhans, Ramona Woitek, Nidhi Goel, Manas Dhir, Deepti Chhabra, Shreshtha Jha, Pallavi Sharma, Vijay Thakur, Simarpreet Singh Chawla, Deepak Gunjan, Jagadeesh Kakarla, and Balasubramanian Raman. Capsule vision 2024 challenge: Multi-class abnormality classification for video capsule endoscopy, 2025. URL https://arxiv.org/abs/2408.04940.

Ali Hatamizadeh. *An Artificial Intelligence Framework for the Automated Segmentation and Quantitative Analysis of Retinal Vasculature*. University of California, Los Angeles, 2020.

Ali Hatamizadeh, Hamid Hosseini, Niraj Patel, Jinseo Choi, Cameron C Pole, Cory M Hoeferlin, Steven D Schwartz, and Demetri Terzopoulos. Ravir: A dataset and methodology for the semantic segmentation and quantitative analysis of retinal arteries and veins in infrared reflectance imaging. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3272–3283, 2022.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Steven A. Hicks, Andrea Storås, Pål Halvorsen, Thomas de Lange, Michael A. Riegler, and Vajira Thambawita. Overview of imageclefmedical 2023 – medical visual question answering for gastrointestinal tract. In *CLEF2023 Working Notes*, CEUR Workshop Proceedings, Thessaloniki, Greece, September 18-21 2023. CEUR-WS.org.

Jindong Hong, Tianjie Chen, Lingjie Luo, Chuanyang Zheng, Ting Xu, Haibao Yu, Jianing Qiu, Qianzhong Chen, Suning Huang, Yan Xu, et al. Benchmarking the thinking mode of multimodal large language models in clinical tasks. *arXiv preprint arXiv:2511.03328*, 2025.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22170–22183, 2024.

Md Nazmul Islam, Mehedi Hasan, Md Kabir Hossain, Md Golam Rabiul Alam, Md Zia Uddin, and Ahmet Soylu. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from ct-radiography. *Scientific Reports*, 12(1):11440, 2022.

Zixi Jia, Jiqiang Liu, Hexiao Li, Qinghua Liu, and Hongbin Gao. Dcot: Dual chain-of-thought prompting for large multimodal models. In *The 16th Asian Conference on Machine Learning (Conference Track)*, 2024.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.

Bai Jieyun and Ou ZhanHong. Pubic symphysis-fetal head segmentation and angle of progression, April 2023. URL https://doi.org/10.5281/zenodo.7851339.

Jakob Nikolas Kather, Frank Gerrit Zöllner, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Cleo-Aron Weis. Collection of textures in colorectal cancer histology, May 2016. URL https://doi.org/10.5281/zenodo.53169.

Ural Koç, Ebru Akçapınar Sezer, Yaşar Alper Özkaya, Yasin Yarbay, Onur Taydaş, Veysel Atilla Ayyıldız, Hüseyin Alper Kızıloğlu, Uğur Kesimal, İmran Çankaya, Muhammed Said Beşler, et al. Artificial intelligence in healthcare competition (teknofest-2021): stroke data set. *The Eurasian journal of medicine*, 54(3):248, 2022.

Zahra Mousavi Kouzehkanan, Sepehr Saghari, Sajad Tavakoli, Peyman Rostami, Mohammadjavad Abaszadeh, Farzaneh Mirzadeh, Esmaeil Shahabi Satlsar, Maryam Gheidishahran, Fatemeh Gorgi, Saeed Mohammadi, et al. A large dataset of white blood cells containing cell locations and types, along with segmented nuclei and cytoplasm. *Scientific reports*, 12(1):1123, 2022.

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.

Wonkyeong Lee, Fabian Wagner, Adrian Galdran, Yongyi Shi, Wenjun Xia, Ge Wang, Xuanqin Mou, Md Atik Ahamed, Abdullah Al Zubaer Imran, Ji Eun Oh, et al. Low-dose computed tomography perceptual image quality assessment. *Medical Image Analysis*, 99:103343, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

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.

Xiaomin Li, Zhou Yu, Zhiwei Zhang, Xupeng Chen, Ziji Zhang, Yingying Zhuang, Narayanan Sadagopan, and Anurag Beniwal. When thinking fails: The pitfalls of reasoning for instruction-following in llms, 2025. URL https://arxiv.org/abs/2505.11423.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, et al. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*, 2025.

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)*, pp. 1650–1654. IEEE, 2021.

Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Zhou, and Zuozhu Liu. Medcot: Medical chain of thought via hierarchical expert. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17371–17389, 2024a.

Shengjie Liu, Chuang Zhu, Feng Xu, Xinyu Jia, Zhongyue Shi, and Mulan Jin. Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1815–1824, June 2022.

Yanzhen Liu, Sutuke Yibulayimu, Yudi Sang, Gang Zhu, Chao Shi, Chendi Liang, Qiyong Cao, Chunpeng Zhao, Xinbao Wu, and Yu Wang. Preoperative fracture reduction planning for image-guided pelvic trauma surgery: A comprehensive pipeline with learning. *Medical Image Analysis*, 102:103506, 2025a. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2025.103506. URL https://www.sciencedirect.com/science/article/pii/S1361841525000544.

Yanzhen Liu, Sutuke Yibulayimu, Gang Zhu, Chao Shi, Chendi Liang, Chunpeng Zhao, Xinbao Wu, Yudi Sang, and Yu Wang. Automatic pelvic fracture segmentation: a deep learning approach and benchmark dataset. *Frontiers in Medicine*, 12:1511487, 2025b.

Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. Chain-of-spot: Interactive reasoning improves large vision-language models. *arXiv preprint arXiv:2403.12966*, 2024b.

Shenghan Lou, Jianxin Ji, Xuan Zhang, Huiying Li, Yang Jiang, Menglei Hua, Kexin Chen, Xiaohan Zheng, Qi Zhang, Peng Han, Lei Cao, and Liuying Wang. Gastric Cancer Histopathology Tissue Image Dataset (GCHTID). 6 2024. doi: 10.6084/m9.figshare.25954813.v1. URL https://figshare.com/articles/dataset/Gastric_Cancer_Histopathology_Tissue_Image_Dataset_GCHTID_/25954813.

Bozhi Luan, Hao Feng, Hong Chen, Yonghui Wang, Wengang Zhou, and Houqiang Li. Textcot: Zoom in for enhanced multimodal text-rich image understanding. *arXiv preprint arXiv:2404.09797*, 2024.

Luca Lumetti, Vittorio Pipoli, Federico Bolelli, Elisa Ficarra, and Costantino Grana. Enhancing Patch-Based Learning for the Segmentation of the Mandibular Canal. *IEEE Access*, pp. 1–12, 2024. ISSN 2169-3536. doi: https://doi.org/10.1109/ACCESS.2024.3408629.

Christian Matek, Sebastian Krappe, Christian Münzenmayer, Torsten Haferlach, and Carsten Marr. An expert-annotated dataset of bone marrow cytology in hematologic malignancies. *The Cancer Imaging Archive*, 2021a.

Christian Matek, Sebastian Krappe, Christian Münzenmayer, Torsten Haferlach, and Carsten Marr. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. *Blood, The Journal of the American Society of Hematology*, 138(20): 1917–1927, 2021b.

Meta AI. Llama-3.2-90b-vision-instruct, September 2024. URL https://huggingface.co/meta-llama/Llama-3.2-90B-Vision-Instruct. Accessed: 2025-05-15.

Aayush Mishra and Karan Thakkar. Stress testing chain-of-thought prompting for large language models. *arXiv preprint arXiv:2309.16621*, 2023.

Agata Momot. Common carotid artery ultrasound images. *Mendeley Data*, 2022.

Anna Montoya, Hasnin, kaggle446, shirzad, Will Cukierski, and yffud. Ultrasound nerve segmentation. https://kaggle.com/competitions/ultrasound-nerve-segmentation, 2016.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pp. 353–367. PMLR, 2023.

Roberto Morelli, Luca Clissa, Roberto Amici, Matteo Cerri, Timna Hitrec, Marco Luppi, Lorenzo Rinaldi, Fabio Squarcio, and Antonio Zoccoli. Automating cell counting in fluorescent microscopy through deep learning with c-resunet. *Scientific Reports*, 11(1):22920, 2021.

Geoff Norman, Meredith Young, and Lee Brooks. Non-analytical models of clinical reasoning: the role of experience. *Medical education*, 41(12):1140–1145, 2007.

OpenAI. Gpt-4 technical report, 2023. URL https://arxiv.org/abs/2303.08774. Accessed: 2025-07-29.

OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024a.

OpenAI. Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms/, 2024b.

OpenAI. Thinking with images. https://openai.com/index/thinking-with-images/, 2024c.

OpenAI. Openai o3-mini. https://openai.com/index/openai-o3-mini/, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid), 2018. URL https://dx.doi.org/10.21227/H25W98.

Bo Qian, Hao Chen, Xiangning Wang, Haoxuan Che, Gitaek Kwon, Jaeyoung Kim, Sungjin Choi, Seoyoung Shin, Felix Krause, Markus Unterdechler, et al. Drac: diabetic retinopathy analysis challenge with ultra-wide optical coherence tomography angiography images. *arXiv preprint arXiv:2304.02389*, 2023.

Gwenolé Quellec, Mathieu Lamard, Pierre-Henri Conze, Pascale Massin, and Béatrice Cochener. Automatic detection of rare pathologies in fundus photographs using few-shot learning. *Medical image analysis*, 61:101660, 2020.

Md Mizanur Rahman. Brain cancer - mri dataset. *Mendeley Data*, 1, 2024.

Tawsifur Rahman, Amith Khandakar, Muhammad Abdul Kadir, Khandaker Rejaul Islam, Khandakar F Islam, Rashid Mazhar, Tahir Hamid, Mohammad Tariqul Islam, Saad Kashem, Zaid Bin Mahbub, et al. Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. *Ieee Access*, 8:191586–191601, 2020.

Tawsifur Rahman, Amith Khandakar, Khandaker Reajul Islam, Md Mohiuddin Soliman, Mohammad Tariqul Islam, et al. Aseptic loose hip implant x-ray database. https://www.kaggle.com/datasets/tawsifurrahman/aseptic-loose-hip-implant-xray-database, 2022.

Netherlands Rotterdam Ophthalmic Institute, Rotterdam Eye Hospital. Justraigs challenge training data set, January 2024. URL https://doi.org/10.5281/zenodo.10035093.

Salman Sajid. Dental condition dataset. https://www.kaggle.com/datasets/salmansajid05/oral-diseases, 2024.

Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13183–13192, 2024.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

FA Sharifullin, DD Dolotova, TG Barmina, SS Petrikov, LS Kokov, GR Ramazanov, YR Blagosklonova, IV Arkhipov, IM Skorobogach, NN Cheremushkin, et al. Creation of a dataset of msct-images and clinical data for acute cerebrovascular events. *Russian Sklifosovsky Journal" Emergency Medical Care"*, 9(2):231–237, 2020.

Chaoyin She, Ruifang Lu, Danni He, Jiayi Lv, Yadan Lin, Meiqing Cheng, Hui Huang, Lida Chen, Wei Wang, and Qinghua Huang. A retrospective systematic study on hierarchical sparse query transformer-assisted ultrasound screening for early hepatocellular carcinoma, 2025. URL https://arxiv.org/abs/2502.03772.

shenggan, Nicolas Chen, cosmicad, and akshaylamba. Bccd: Blood cell count and detection, 2018. URL https://github.com/Shenggan/BCCD_Dataset.

Osamah Taher and Kasım Özacar. Hecapsnet: An enhanced capsule network for automated heel disease diagnosis using lateral foot x-ray images. *International Journal of Imaging Systems and Technology*, 34(3):e23084, 2024.

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

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey, 2025. URL https://arxiv.org/abs/2503.12605.

Alexander Wong, Hayden Gunraj, Vignesh Sivan, and Masoom A. Haider. Synthetic correlated diffusion imaging hyperintensity delineates clinically significant prostate cancer. *Scientific Reports*, 12(3376), 2022. doi: 10.1038/s41598-022-06872-7.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *arXiv preprint arXiv:2308.02463*, 2023.

Peiran Wu, Che Liu, Canyu Chen, Jun Li, Cosmin I Bercea, and Rossella Arcucci. Fmbench: Benchmarking fairness in multimodal large language models on medical tasks. *arXiv preprint arXiv:2410.01089*, 2024.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024a.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step, 2024b. URL https://arxiv.org/abs/2411.10440.

Pusheng Xu, Xiaolan Chen, Ziwei Zhao, and Danli Shi. Evaluation of a digital ophthalmologist app built by gpt4-v (ision). *medRxiv*, pp. 2023–11, 2023.

Qianqi Yan, Yue Fan, Hongquan Li, Shan Jiang, Yang Zhao, Xinze Guan, Ching-Chen Kuo, and Xin Eric Wang. Multimodal inconsistency reasoning (mmir): A new benchmark for multimodal reasoning models. *arXiv preprint arXiv:2502.16033*, 2025.

Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37: 94327–94427, 2024.

Xu Yiming, Zheng Bowen, Liu Xiaohong, Wu Tao, Ju Jinxiu, Wang Shijie, Lian Yufan, Zhang Hongjun, Liang Tong, Sang Ye, Jiang Rui, Wang Guangyu, Ren Jie, and Chen Ting. Annotated ultrasound liver images, November 2022. URL https://doi.org/10.5281/zenodo.7272660.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

APPENDIX

This supplementary material provides more detailed information about M3CoTBench. The content of each appendix is summarized as follows:

- **Appendix A** Provides a detailed description of how large language models are applied in this work. This includes their use in assisting writing, guiding dataset construction, supporting annotation processes, and contributing to model evaluation.
- **Appendix B** Offers comprehensive information about the dataset used in this study, including the sources of the data, the diseases and abnormalities covered, the distribution of image resolutions, detailed task specifications in the benchmark, and descriptions of Chain-of-Thought (CoT) annotations.
- **Appendix C** Provides an in-depth explanation of the evaluation methodology, including the metrics used, the design of prompts, and additional clarifications on how model performance is measured and interpreted.
- **Appendix D** Presents supplementary experimental results that complement the main paper, along with illustrative case studies that demonstrate model behavior and practical outcomes in various scenarios.
- **Appendix E** Discusses the known limitations of this study, including potential weaknesses in the methodology, dataset coverage, and model generalizability, providing a balanced view of the research.
- **Appendix F** Highlights potential societal implications of this work, considering both beneficial applications and possible risks, and reflecting on the broader impact of deploying such models in real-world scenarios.

A THE USE OF LARGE LANGUAGE MODELS

We use large language models solely for polishing our writing, and we have conducted a careful check, taking full responsibility for all content in this work. In addition, LLMs and MLLMs were also used in the construction of the dataset and the evaluation of models, and the specific usage has been described in detail in the main text.

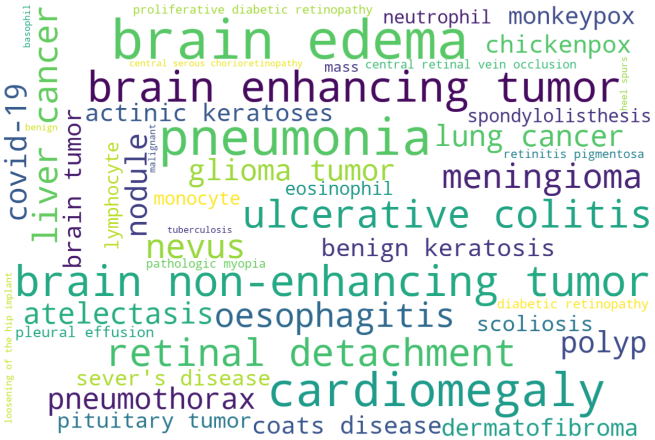


Figure A1: **Word cloud for abnormality and diseases included in M3CoTBench** The word cloud below visualizes the frequency and variety of these labels, highlighting the spectrum of diagnostic conclusions and imaging findings represented.

B More Details about the Dataset

B.1 Source Dataset Information

Images in the M3CoTBench dataset are collected from 55 publicly available datasets, offering a highly diverse and representative foundation for training and evaluating multi-modal medical reasoning models. Its comprehensive coverage across modalities, anatomies, time periods, and geographic sources ensures broad applicability and robustness in real-world clinical scenarios. The detailed information can be seen in Table A1.

B.2 Diseases and Abnormalities

This dataset contains a wide range of diseases and abnormalities. A word cloud illustrating their distribution is shown in Figure A1.

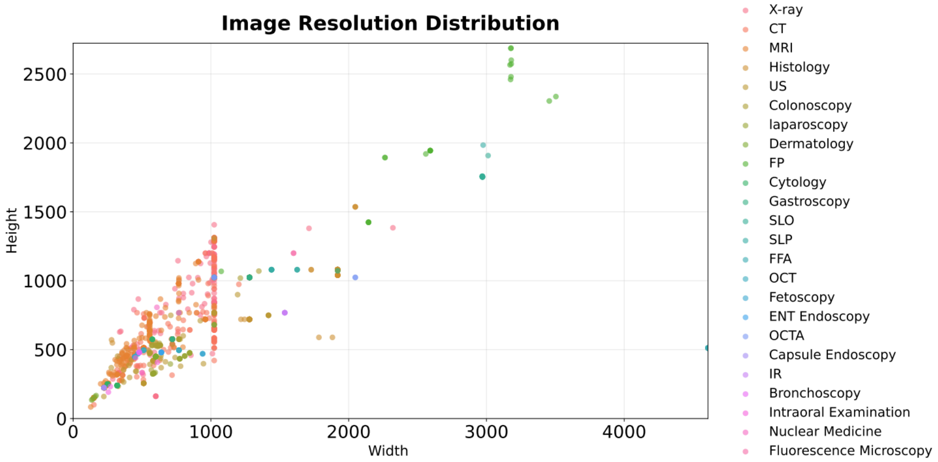


Figure A2: **Image resolution distribution in M3CoTBench.** Most images are concentrated below a width of 1200 and a height of 1500, though some exhibit higher resolutions.

B.3 Image Resolution Distribution

For the images, we retained their original sizes as provided in the source datasets, without applying any additional compression or resizing. Some images may have been preprocessed in their original datasets. However, for tasks such as entity linking, grading, and image quality comparison, we concatenate two images side by side, which results in increased image width. The resolution distribution information can be seen in Figure A2.

- **Diversity in examination types:** The dataset covers 24 imaging modalities and examination methods, which can be grouped into six major categories: ophthalmic imaging, radiology, endoscopy, microscopy, ultrasound-based examinations, and surface-level inspections. These include slit lamp photography (SLP), fundus photography (FP), optical coherence tomography (OCT), optical coherence tomography angiography (OCTA), scanning laser ophthalmoscopy (SLO), fundus fluorescein angiography (FFA), X-ray, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound (US), infrared reflectance (IR), nuclear medicine, fetoscopy, laparoscopy, colonoscopy, gastroscopy, capsule endoscopy, bronchoscopy, ENT endoscopy, cytology, fluorescence microscopy, dermoscopy, and intraoral examination.
- **Diversity in anatomical regions:** The datasets encompass a broad spectrum of anatomical regions, including but not limited to the eye, skin, chest (lungs and heart), brain, abdomen (liver, kidney, stomach, etc.), oral cavity, uterus and fetal environment, breast, vertebrae, hip, knee, foot, blood, and bone marrow. This anatomical diversity supports the evaluation of models' capability across different clinical tasks and organ systems.
- **Diversity in publication years:** The included datasets were published across a wide temporal range, from earlier benchmarks to very recent contributions. This time span captures the evolution

of imaging quality, annotation practices, and diagnostic standards, making the dataset suitable for both historical benchmarking and future-proof model evaluation.

- **Geographic diversity:** The data sources originate from over a dozen countries and regions, reflecting a variety of healthcare environments, population demographics, and medical imaging protocols. This geographic diversity enhances the robustness, fairness, and real-world applicability of models trained on the dataset, particularly in cross-domain or multi-institutional settings. The geographic distribution of data sources is illustrated in FigureA3.

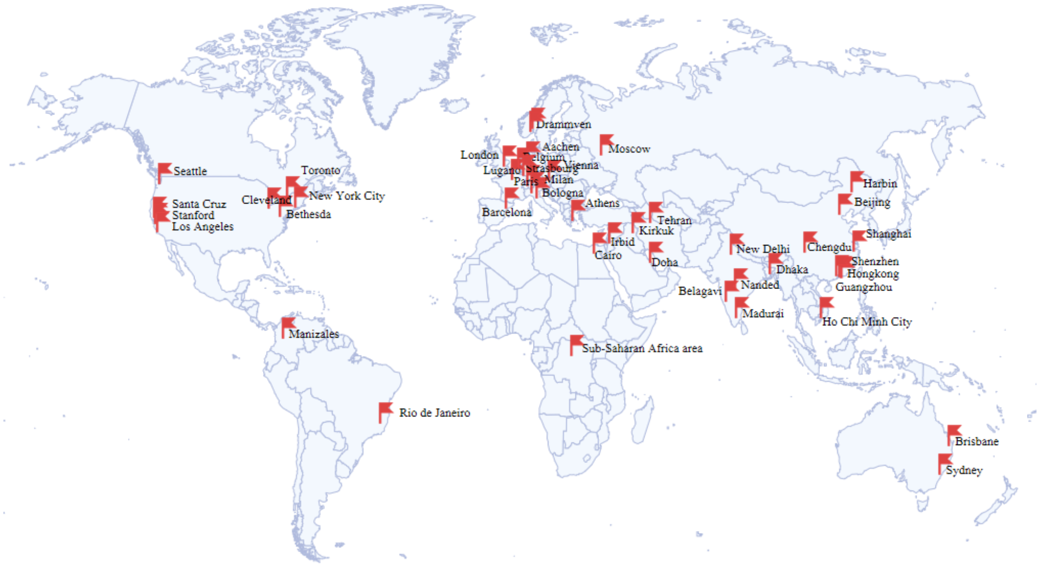


Figure A3: Geographic distribution of data sources in the dataset. Red flags indicate the locations of contributing hospitals or institutions, where applicable. Due to the complex and varied origins of some datasets, exact source locations may not always be clearly identifiable.

B.4 Detailed Introduction to Tasks

The benchmark encompasses a diverse range of tasks that mirror real-world clinical challenges in medical visual-language reasoning. These tasks are designed to evaluate not only a model's ability to recognize and classify visual information, but also its capacity to comprehend spatial, procedural, and diagnostic contexts. Broadly, the tasks can be grouped into two conceptual levels: **Perceptual-level tasks** focus on low- to mid-level visual understanding, such as identifying image modality, recognizing anatomical structures, or assessing image quality. These tasks primarily test the model's capability to extract and interpret observable features from the image. **Knowledge-based reasoning tasks**, on the other hand, require integration of visual features with clinical knowledge, commonsense reasoning, or multi-step inference. These include complex tasks such as diagnosing diseases, predicting disease progression, grading severity, planning clinical actions, or identifying causal relationships.

- **Modality / Examination Types:** Understanding and recognizing the imaging modality involved, such as CT, MRI, X-ray, or OCT, demonstrates the model's awareness of different diagnostic techniques and their clinical contexts.
- **Image Quality Assessment:** Evaluating whether an image is diagnostically adequate, and comparing the relative quality between multiple images when necessary. This reflects the model's ability to judge image usability in clinical practice.
- **Recognition:** General visual recognition tasks, including identifying anatomical structures, tissues, or medical devices, without explicit spatial reference.
- **Referring Recognition:** Region-specific identification tasks where the model must recognize or interpret a particular area in the image based on the question or accompanying text.

Table A1: Data sources of different modalities in M3CoTBench

Dataset	**Anatomical Region**	**Modality / Examination Type**
OphthalVQA (Xu et al., 2023)	Eye	SLP, FP, OCT, US, SLO, FFA
IDRiD (Porwal et al., 2018)	Eye	FP
JustRAIGS (Rotterdam Ophthalmic Institute, 2024)	Eye	FP
RIADD (Quellec et al., 2020)	Eye	FP
DRAC 2022 (Qian et al., 2023)	Eye	OCTA
RAVIR (Hatamizadeh et al., 2022; Hatamizadeh, 2020)	Eye	IR
ISIC 2018 (Codella et al., 2019)	Skin	Dermoscopy
HAM10000 (Tschandl et al., 2018)	Skin	Dermoscopy
MSLD v2.0 (Ali et al., 2024)	Skin	Dermoscopy
VQA-RAD (Lau et al., 2018)	Chest, Abdomen, Brain	X-ray, CT, MRI, Nuclear Medicine
VQA-Med-2019 (Ben Abacha et al., 2019)	Chest, Abdomen, Brain	X-ray, CT, MRI, US
SLAKE (Liu et al., 2021)	Chest, Abdomen, Brain	X-ray, CT, MRI
Chest XR COVID-19 (Akhloufi & Chetoui, 2021)	Chest (Lung)	X-ray
TB Chest X-ray (Rahman et al., 2020)	Chest (Lung)	X-ray
Heel Bone (Taher & Özacar, 2024)	Foot	X-ray
Digital Knee X-ray (Gornale & Patravali, 2020)	Knee	X-ray
Vertebrae X-ray (Fraiwan et al., 2022)	Vertebrae	X-ray
Hip Implant X-ray (Rahman et al., 2022)	Shoulder	X-ray
CT Kidney (Islam et al., 2022)	Kidney	CT
COVID-19 Lung CT (cov, 2020)	Lung	CT
Brain Stroke CT (Koç et al., 2022)	Brain	CT
LDCTIQAC 2023 (Lee et al., 2025)	Abdomen	CT
MSCT-Image Dataset (Sharifullin et al., 2020)	Brain	CT
PENGWIN (Liu et al., 2025a;b)	Pelvis	CT
ToothFairy (Bolelli et al., 2024; Lumetti et al., 2024; Cipriano et al., 2022)	Oral Cavity	CT
Brain Tumor (Bhuvaji et al., 2020)	Brain	MRI
LGG Segmentation (Buda et al., 2019)	Brain	MRI
Brain Cancer MRI (Rahman, 2024)	Brain	MRI
BRATS-SSA (Adewole et al., 2023)	Brain	MRI
Cancer-Net PCa-Data (Wong et al., 2022; Gunraj et al., 2023)	Prostate	MRI
SIMON MRI (Duchesne et al., 2019)	Brain	MRI
BUSI (Al-Dhabyani et al., 2020)	Breast	US

Table A1 (continued): Data sources of different modalities in M3CoTBench

Dataset	**Anatomical Region**	**Modality / Examination Type**
FH-PS-AOP (Jieyun & Zhan-Hong, 2023)	Fetal	US
Nerve Segmentation (Montoya et al., 2016)	Neck	US
Carotid Artery (Momot, 2022)	Neck	US
Liver-US (She et al., 2025)	Liver	US
Annotated Liver US Dataset (Yiming et al., 2022)	Liver	US
BUS-BRA (Gómez-Flores et al., 2024)	Breast	US
Quilt-VQA (Seyfioglu et al., 2024)	Multi-regions	Histology
BCI (Liu et al., 2022)	Breast	Histology
Colorectal Histology MNIST (Kather et al., 2016)	Colon and Rectum	Histology
GCHTID (Lou et al., 2024)	Stomach	Histology
Dental Condition Dataset (Sajid, 2024)	Oral Cavity	Intraoral Examination
BACH (Aresta et al., 2019)	Breast	Histology
CMIA Histological Slides	Lung, Breast	Histology
Fluorescent Neuronal Cells (Morelli et al., 2021)	Brain	Fluorescent Microscopy
BCCD (shenggan et al., 2018)	Blood	Cytology
Raabin-WBC (Kouzehkanan et al., 2022)	Blood	Cytology
BMC (Matek et al., 2021a;b)	Bone Marrow	Cytology
EndoVis-17-VLQA (Allan et al., 2019)	Abdomen	Laparoscopy
m2cai16-tool (Twinanda et al., 2016)	Abdomen	Laparoscopy
ImageCLEFmed MEDVQA-GI (Hicks et al., 2023)	Gastrointestinal Tract	Colonoscopy, Gastroscopy
Bronchoscopy Dataset (Deng et al., 2023)	Airway Tract	Bronchoscopy
Capsule Vision 2024 (Handa et al., 2025)	Gastrointestinal Tract	Capsule Endoscopy
ENTRep Challenge 2025 (ENT, 2025)	Ear, Nose, Throat	ENT Endoscopy
FetReg (Bano et al., 2021)	Uterus / Fetal Environment	Fetoscopy
Fetoscopy Placenta Data (Bano et al., 2020)	Uterus / Fetal Environment	Fetoscopy

- **Counting:** Quantifying specific elements in an image, such as surgical tools, lesions, polyps, or cells, often requiring precise object detection and differentiation.
- **Localization:** Identifying the spatial location of regions of interest, such as lesions, organs, or abnormal structures, testing the model's understanding of spatial relations and context.
- **Diagnosis:** Inferring the presence of abnormalities, diseases, or clinical conditions based on image and text input; this is the most common and clinically important task category.
- **Grading:** Assessing the severity or stage of a medical condition, such as cancer staging or diabetic retinopathy levels, requires a nuanced interpretation of visual cues.
- **Symptom Identification:** Recognizing observable clinical signs or inferring underlying symptoms based on the visual features of the image and contextual cues.
- **Clinical Action Planning:** Making decisions about the next steps in patient care, such as recommending further examinations, procedures, or treatment options, demonstrating clinical reasoning ability.
- **Prediction:** Estimating future disease progression, risks of complications, or expected outcomes, often involving multi-modal reasoning over image and text inputs.
- **Functional Understanding:** Interpreting the physiological function of organs, the intended use of medical instruments, or the purpose of surgical actions, integrating procedural and anatomical knowledge.
- **Causal Reasoning:** Identifying the cause or etiology of a symptom or condition, requiring the model to reason about potential underlying mechanisms or prior events.

B.5 CoT Annotation

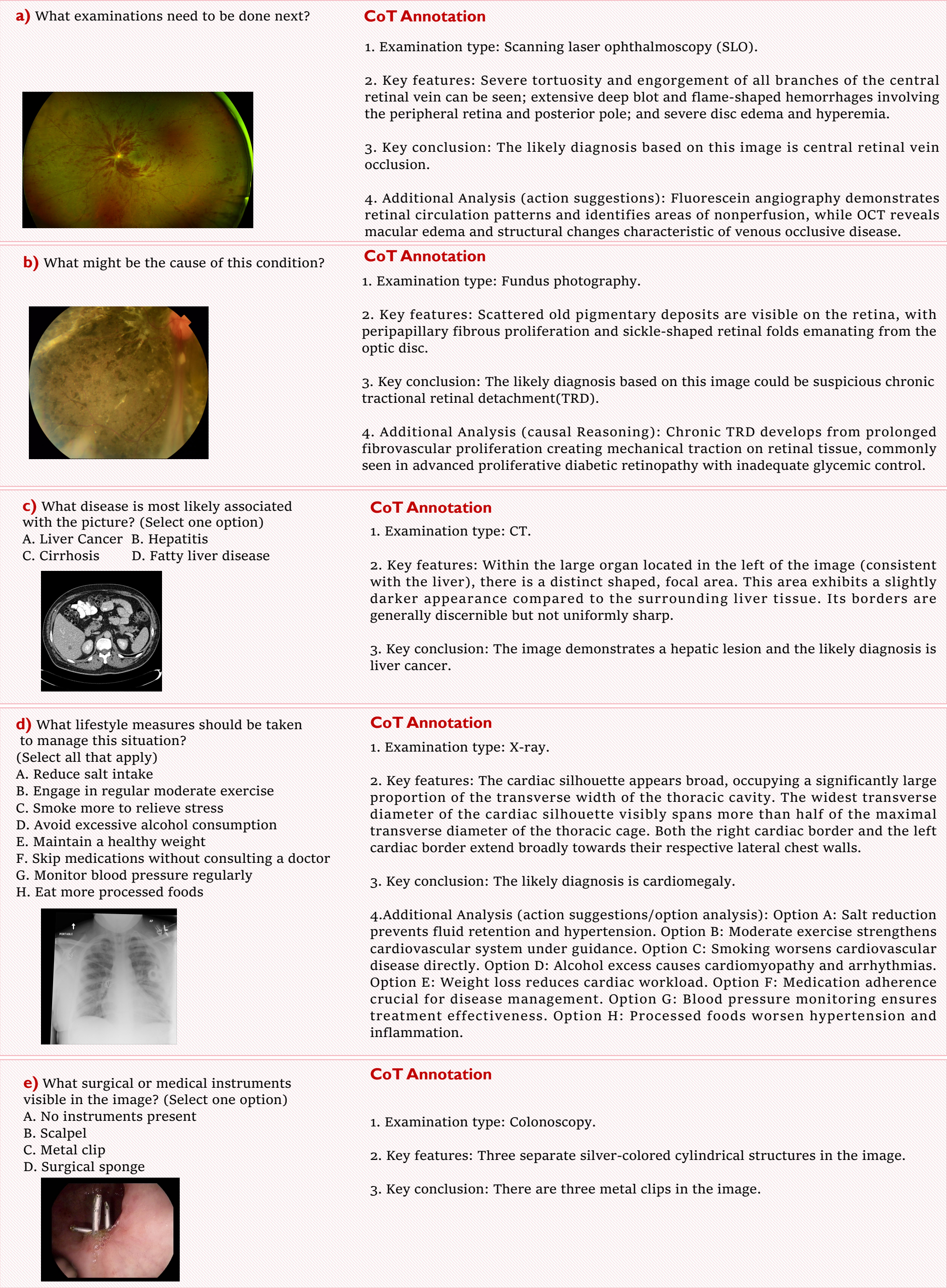
The CoT annotations are collaboratively generated by medical experts and MLLMs, generally following a four-part structure: {examination type, key feature, key conclusion, additional analysis}. This approach aligns closely with clinical reasoning patterns used by physicians, who often begin by identifying the type of examination or modality, observing key findings, deriving conclusions, and, when necessary, conducting further interpretation or differential diagnosis. The length and structure of CoT vary depending on the task. For tasks such as recognition, diagnosis, and grading, a three-step format,{examination type, key feature, key conclusion}, is generally sufficient. In contrast, more complex tasks like treatment planning, causal reasoning, symptom analysis, prognostic prediction, or functional interpretation often require a four-step annotation to capture the depth of reasoning. When it comes to identifying the imaging modality, CoT length depends on the nature of the question. For example, in general tasks, it may not be necessary to analyze image features to identify the modality explicitly. However, in questions specifically targeting modality identification, CoT annotations typically include two steps, focusing on characteristic visual clues about the imaging technique used. Notably, during examination modality statistics, some subtypes are grouped into broader categories. However, in CoT annotations, these modalities are often distinguished more finely. For example, IHC and HE are treated separately, as are MRI T1-weighted and T2-weighted images. Examples of CoT annotation are shown in Figure A4, Figure A5, and Figure A6.

C Supplementary to Evaluation Processes

C.1 Examples of Path Similarity in Reasoning Consistency Evaluation

To evaluate the structural stability of reasoning in multi-step tasks, we introduce a **path consistency** metric that measures the similarity of reasoning paths across instances of the same task type. Unlike traditional metrics that treat reasoning steps as unordered, this metric accounts for the sequential structure by comparing generated paths to a canonical reference using the normalized longest common subsequence (LCS). The final score, averaged over all 13 tasks, reflects the model's ability to follow consistent, interpretable reasoning patterns, a key property for transparency and trust in medical decision-making. Here are some examples to show the specific calculation method:

Consider the following reasoning paths, where each element is one of {*modality*, *feature*, *diagnosis*, *treatment*}, representing a progression from identifying the imaging type, describing visual findings, inferring clinical conditions, to suggesting appropriate medical interventions.



a) What examinations need to be done next?

CoT Annotation

1. Examination type: Scanning laser ophthalmoscopy (SLO).

2. Key features: Severe tortuosity and engorgement of all branches of the central retinal vein can be seen; extensive deep blot and flame-shaped hemorrhages involving the peripheral retina and posterior pole; and severe disc edema and hyperemia.

3. Key conclusion: The likely diagnosis based on this image is central retinal vein occlusion.

4. Additional Analysis (action suggestions): Fluorescein angiography demonstrates retinal circulation patterns and identifies areas of nonperfusion, while OCT reveals macular edema and structural changes characteristic of venous occlusive disease.

b) What might be the cause of this condition?

CoT Annotation

1. Examination type: Fundus photography.

2. Key features: Scattered old pigmentary deposits are visible on the retina, with peripapillary fibrous proliferation and sickle-shaped retinal folds emanating from the optic disc.

3. Key conclusion: The likely diagnosis based on this image could be suspicious chronic tractional retinal detachment(TRD).

4. Additional Analysis (causal Reasoning): Chronic TRD develops from prolonged fibrovascular proliferation creating mechanical traction on retinal tissue, commonly seen in advanced proliferative diabetic retinopathy with inadequate glycemic control.

c) What disease is most likely associated with the picture? (Select one option)
A. Liver Cancer B. Hepatitis
C. Cirrhosis D. Fatty liver disease

CoT Annotation

1. Examination type: CT.

2. Key features: Within the large organ located in the left of the image (consistent with the liver), there is a distinct shaped, focal area. This area exhibits a slightly darker appearance compared to the surrounding liver tissue. Its borders are generally discernible but not uniformly sharp.

3. Key conclusion: The image demonstrates a hepatic lesion and the likely diagnosis is liver cancer.

d) What lifestyle measures should be taken to manage this situation?
(Select all that apply)
A. Reduce salt intake
B. Engage in regular moderate exercise
C. Smoke more to relieve stress
D. Avoid excessive alcohol consumption
E. Maintain a healthy weight
F. Skip medications without consulting a doctor
G. Monitor blood pressure regularly
H. Eat more processed foods

CoT Annotation

1. Examination type: X-ray.

2. Key features: The cardiac silhouette appears broad, occupying a significantly large proportion of the transverse width of the thoracic cavity. The widest transverse diameter of the cardiac silhouette visibly spans more than half of the maximal transverse diameter of the thoracic cage. Both the right cardiac border and the left cardiac border extend broadly towards their respective lateral chest walls.

3. Key conclusion: The likely diagnosis is cardiomegaly.

4.Additional Analysis (action suggestions/option analysis): Option A: Salt reduction prevents fluid retention and hypertension. Option B: Moderate exercise strengthens cardiovascular system under guidance. Option C: Smoking worsens cardiovascular disease directly. Option D: Alcohol excess causes cardiomyopathy and arrhythmias. Option E: Weight loss reduces cardiac workload. Option F: Medication adherence crucial for disease management. Option G: Blood pressure monitoring ensures treatment effectiveness. Option H: Processed foods worsen hypertension and inflammation.

e) What surgical or medical instruments visible in the image? (Select one option)
A. No instruments present
B. Scalpel
C. Metal clip
D. Surgical sponge

CoT Annotation

1. Examination type: Colonoscopy.

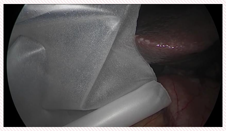
2. Key features: Three separate silver-colored cylindrical structures in the image.

3. Key conclusion: There are three metal clips in the image.

Figure A4: **Examples of CoT annotations with corresponding images and questions in M3CoTBench (1).** Different types of questions are annotated with different lengths of CoT steps. For example, diagnostic **(c)** and recognition **(e)** questions involve three annotation steps, while action-planning **(a, d)** and causal analysis **(b)** questions are annotated with four steps.

a) What are the primary functions of the instrument in this image? (Select all that apply)

A. Retrieval of resected tissue or polyps
B. Coagulation of bleeding vessels
C. Prevention of contamination or spillage during specimen removal
D. Inflation of the abdominal cavity
E. Visualization enhancement
F. Safe extraction of specimens through trocars
G. Biopsy sampling
H. Closure of mucosal defects



CoT Annotation

1. Examination type: Laparoscopy.

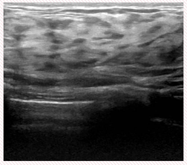
2. Key features: The visible tool is light gray to white in color with a matte, slightly textured surface. It appears to be made of a thin, flexible material, exhibiting multiple folds and drapes. A section of the material is visible in a rolled or gathered configuration.

3. Key conclusion: The tool shown in the image is specimen bag.

4. Additional Analysis (functions): Option A: Designed to contain and remove resected tissues from body cavity. Option B: Plastic bag lacks energy source for coagulation. Option C: Prevents spillage of infectious contents and malignant cell seeding. Option D: Insufflation achieved through specialized trocar, not specimen bag. Option E: Visualization is endoscope function, not retrieval bag. Option F: Contains specimen allowing extraction through small trocar safely. Option G: Used for retrieval after resection, not for taking samples. Option H: Closure performed with clips or sutures, not retrieval bags.

b) What does this image most likely represent? (Select one option)

A. Normal B. Benign
C. Malignant D. None of above



CoT Annotation

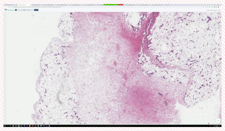
1. Examination type: Ultrasound.

2. Key features:The image displays a heterogeneous, mottled gray echotexture in the upper portion. Below this, there are distinct, thin, parallel hyperechoic (bright) linear structures interspaced with hypoechoic (darker gray) regions. These linear structures are smooth and appear continuous. The overall echotexture appears organized and consistent throughout the depicted area. There is an absence of distinct, irregularly shaped focal anechoic (black) or intensely hyperechoic (white) lesions, or areas of distorted architecture.

3. Key conclusion: The diagnosis conclusion is Normal.

c) What kind of pathological process does the image most likely suggest? (Select one option)

A. Inflammatory edema
B. Hemangioma formation
C. Lymphedema-like process
D. Necrotizing lesion



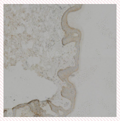
CoT Annotation

1. Examination type: Section stained with hematoxylin and eosin (H&E).

2. Key features: The image displays extensive areas of packed, eosinophilic (pink-stained) fibrous tissue. Within both the pink fibrous tissue and the adjacent adipose tissue, irregular, clear, and empty or very pale-staining spaces of varying sizes are observable. The white, vacuolated fat cells within the adipose tissue appear separated by pale, amorphous material or thin pink septa.

3. Key conclusion: The image suggests Lymphedema-like process.

d) True or False：The staining method shown in the image is hematoxylin and eosin (H&E) staining.

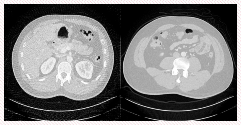


CoT Annotation

1. Examination type: Immunohistochemical (IHC) staining.

2. Key features: The image shows scattered brown reaction product (likely DAB) over a very lightly counter-stained background.

e) True or False: The image on the left is of higher quality than the one on the right.



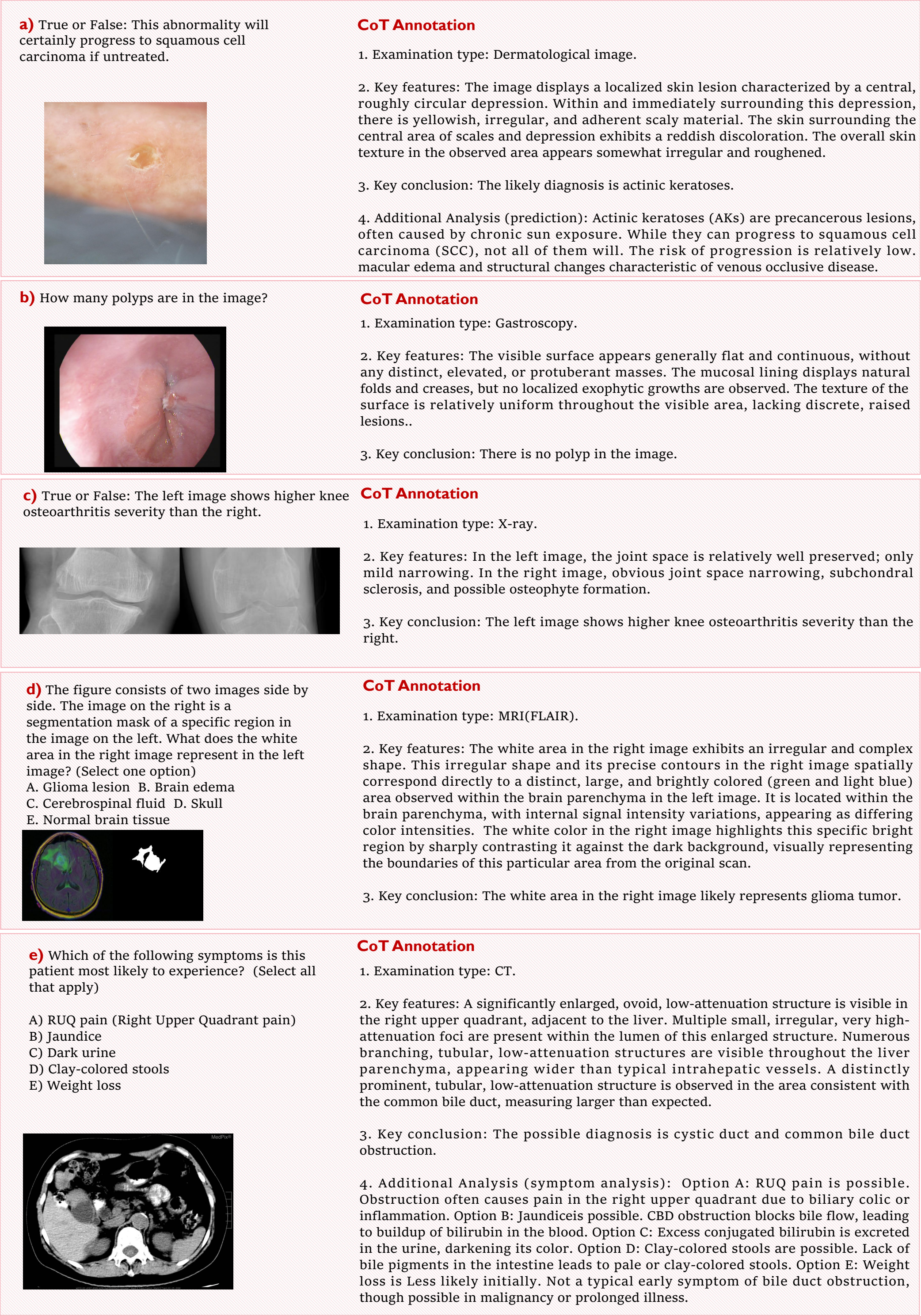
CoT Annotation

1. Examination type: CT.

2. Key features:The image on the right has a much smoother appearance with significantly less grainy noise compared to the left.The image on the right shows fewer streaking artifacts, especially visible around the body contour, which are prominent in the left image. Better Soft Tissue Delineation : Structures and boundaries within the soft tissues (e.g., bowel loops, fat planes) are more clearly defined and have better contrast on the right.

3. The image on the left is of higher quality than the one on the right.

Figure A5: **Examples of CoT annotations with corresponding images and questions in M3CoTBench (2).** Different types of questions are annotated with different lengths of CoT steps. For example, diagnostic **(b, c)** and image quality **(d)** questions involve three annotation steps, function understanding **(a)** question is annotated with four steps, and examination type **(d)** is annotated with two steps as the modality is the conclusion itself.



a) True or False: This abnormality will certainly progress to squamous cell carcinoma if untreated.

CoT Annotation

1. Examination type: Dermatological image.

2. Key features: The image displays a localized skin lesion characterized by a central, roughly circular depression. Within and immediately surrounding this depression, there is yellowish, irregular, and adherent scaly material. The skin surrounding the central area of scales and depression exhibits a reddish discoloration. The overall skin texture in the observed area appears somewhat irregular and roughened.

3. Key conclusion: The likely diagnosis is actinic keratoses.

4. Additional Analysis (prediction): Actinic keratoses (AKs) are precancerous lesions, often caused by chronic sun exposure. While they can progress to squamous cell carcinoma (SCC), not all of them will. The risk of progression is relatively low. macular edema and structural changes characteristic of venous occlusive disease.

b) How many polyps are in the image?

CoT Annotation

1. Examination type: Gastroscopy.

2. Key features: The visible surface appears generally flat and continuous, without any distinct, elevated, or protuberant masses. The mucosal lining displays natural folds and creases, but no localized exophytic growths are observed. The texture of the surface is relatively uniform throughout the visible area, lacking discrete, raised lesions..

3. Key conclusion: There is no polyp in the image.

c) True or False: The left image shows higher knee osteoarthritis severity than the right.

CoT Annotation

1. Examination type: X-ray.

2. Key features: In the left image, the joint space is relatively well preserved; only mild narrowing. In the right image, obvious joint space narrowing, subchondral sclerosis, and possible osteophyte formation.

3. Key conclusion: The left image shows higher knee osteoarthritis severity than the right.

d) The figure consists of two images side by side. The image on the right is a segmentation mask of a specific region in the image on the left. What does the white area in the right image represent in the left image? (Select one option)
A. Glioma lesion B. Brain edema
C. Cerebrospinal fluid D. Skull
E. Normal brain tissue

CoT Annotation

1. Examination type: MRI(FLAIR).

2. Key features: The white area in the right image exhibits an irregular and complex shape. This irregular shape and its precise contours in the right image spatially correspond directly to a distinct, large, and brightly colored (green and light blue) area observed within the brain parenchyma in the left image. It is located within the brain parenchyma, with internal signal intensity variations, appearing as differing color intensities. The white color in the right image highlights this specific bright region by sharply contrasting it against the dark background, visually representing the boundaries of this particular area from the original scan.

3. Key conclusion: The white area in the right image likely represents glioma tumor.

e) Which of the following symptoms is this patient most likely to experience? (Select all that apply)

A) RUQ pain (Right Upper Quadrant pain)
B) Jaundice
C) Dark urine
D) Clay-colored stools
E) Weight loss

CoT Annotation

1. Examination type: CT.

2. Key features: A significantly enlarged, ovoid, low-attenuation structure is visible in the right upper quadrant, adjacent to the liver. Multiple small, irregular, very high-attenuation foci are present within the lumen of this enlarged structure. Numerous branching, tubular, low-attenuation structures are visible throughout the liver parenchyma, appearing wider than typical intrahepatic vessels. A distinctly prominent, tubular, low-attenuation structure is observed in the area consistent with the common bile duct, measuring larger than expected.

3. Key conclusion: The possible diagnosis is cystic duct and common bile duct obstruction.

4. Additional Analysis (symptom analysis): Option A: RUQ pain is possible. Obstruction often causes pain in the right upper quadrant due to biliary colic or inflammation. Option B: Jaundiceis possible. CBD obstruction blocks bile flow, leading to buildup of bilirubin in the blood. Option C: Excess conjugated bilirubin is excreted in the urine, darkening its color. Option D: Clay-colored stools are possible. Lack of bile pigments in the intestine leads to pale or clay-colored stools. Option E: Weight loss is Less likely initially. Not a typical early symptom of bile duct obstruction, though possible in malignancy or prolonged illness.

Figure A6: **Examples of CoT annotations with corresponding images and questions in M3CoTBench (3).** Different types of questions are annotated with different lengths of CoT steps. For example, counting **(b)**, grading **(c)** and referring recognition **(d)** questions involve three annotation steps, and prediction **(a)** and symptom **(e)** questions are annotated with four steps.

- **Example 1:** P_1 = [modality, feature, diagnosis], P_2 = [feature, modality, diagnosis]. Then the LCS is [modality, diagnosis] and [feature, diagnosis]. $\left|\mathrm{LCS}(P_1, P_2)\right| = 2$, thus

$$\mathrm{sim}((P_1, P_2) \;=\; \frac{2}{\max(3, 3)} \;=\; \tfrac{2}{3} \approx 0.67. \tag{A1}$$

- **Example 2:** P_1 = [modality, diagnosis, treatment], P_2 = [modality, feature, diagnosis, treatment]. Then the LCS is [modality, diagnosis, treatment] $\left|\mathrm{LCS}(P_1, P_2)\right| = 3$, thus

$$\mathrm{sim}((P_1, P_2) \;=\; \frac{3}{\max(3, 4)} \;=\; \tfrac{3}{4} = 0.75. \tag{A2}$$

- **Example 3:** P_1 = [modality, feature, treatment] , P_2 = [modality, feature, diagnosis, treatment]. Then the LCS is [modality, diagnosis, treatment] The $\left|\mathrm{LCS}(P_1, P_2)\right| = 3$, thus

$$\mathrm{sim}((P_1, P_2) \;=\; \frac{3}{\max(3, 4)} \;=\; \tfrac{3}{4} = 0.75. \tag{A3}$$

- **Example 4:**

  P_1 = [feature, modality, diagnosis, treatment] , P_2 = [modality, feature, diagnosis, treatment]. Then the LCS is [modality, diagnosis, treatment] and [feature, diagnosis, treatment] The $\left|\mathrm{LCS}(P_1, P_2)\right| = 3$, thus

$$\mathrm{sim}((P_1, P_2) \;=\; \frac{3}{\max(3, 4)} \;=\; \tfrac{3}{4} = 0.75. \tag{A4}$$

C.2 Evaluation Prompts

During evaluation, we use GPT-4o to assess the correctness of each step. Since the feature description and additional analysis parts of the CoT annotations are relatively subjective, with multiple valid expressions for the same meaning, we adopt more lenient instructions for these components. In contrast, we apply stricter criteria to the examination modality and key conclusion steps.

C.2.1 Evaluation prompts for answer accuracy

The prompt for calculating accuracy for both direct outputs and CoT outputs is shown below:

Prompt for calculating accuracy for both direct outputs and CoT outputs

You are a medical evaluation expert:

#Your tasks:
1. From the model's prediction below, **extract the final answer only** (ignore reasoning, explanations, or intermediate answers).
2. Judge whether this extracted final answer matches the provided ground-truth answer.
#Type instruction:
Return ONLY a JSON object with the EXACT format below (no extra text):

```
[
{{
  "match": true or false,
  "final_answer": "the extracted final
  answer text"
}}
```

```
Inputs:

Question:
{question}

Ground-truth Answer:
{answer}

Model's Prediction:
{prediction}
```

C.2.2 Evaluation prompts for precision calculation

The prompt for precision calculation is:

Prompt for calculating precision for CoT outputs

Given a solution with multiple reasoning steps for an image-based problem, reformat it into well-structured steps and evaluate their correctness:

Step 1: Reformatting the Solution
Convert the unstructured solution into distinct reasoning steps while:

- Preserving all original content and order.
- Not adding new interpretations.
- Not omitting any steps.

# Step Types

1. Image Modality or Exam Method
   - Describes the imaging type or procedure used (e.g., CT, MRI).
   - Focuses on technical aspects without interpretation.
2. Key Image Feature Analysis
   - Pure visual observations.
   - Describes visible structures or abnormalities in the image.
   - Pure observation without inference.
3. Identification, Localization, or Diagnostic Conclusion
   - Provides specific findings or diagnosis based on image features.
   - Includes reasoning and clinical conclusions.

  - Classification conclusion for cells or organs.

4. Knowledge-Based / Differential / Exploratory Analysis
   - Includes disease progression prediction, organ/cell/instrument function, treatment or further examination suggestions, cause analysis of disease or abnormalities, other medical knowledge, and step-by-step analysis of multiple-choice options.

# Step Requirements

- Each step must be atomic (one conclusion per step)
- No content duplication across steps
- Initial analysis counts as background information
- Final answer determination counts as logical inference

Step 2: Evaluating Correctness
Evaluate each step against:
# Ground Truth Matching

- For modality or examination types: Must strictly correspond to ground truth; different wording allowed if meaning is equivalent.
- For image feature description: Lenient matching, largely overlap and similar meaning with ground truth are fully accepted as correct, as long as there is no contradiction.
- For key conclusions: Should strictly correspond to ground truth; different wording allowed if meaning is matched or entailed.
- For additional analysis: Lenient matching, largely overlap and similar meaning with ground truth are fully accepted as correct, as long as there is no contradiction.

# Reasonableness Check

- Premises must not contradict any ground truth or correct answer.
- Logic is valid.
- Conclusion must not contradict any ground truth.
- Conclusion must support or be neutral to the correct answer.

# Judgement Categories

- Match: Aligns with ground truth.
- Reasonable: Valid but not in ground truth.
- Wrong: Invalid or contradictory.
- N/A: For background information steps.

# Output Requirements

1. The output format MUST be in valid `JSON` format without ANY other content.
2. For highly repetitive patterns, output it as a single step.

Here is the JSON output format:

```
[
  {
    "step_type": "image description | logical inference
    | background information",
    "premise": "Supporting evidence
    (required only for logical inference)",
    "conclusion": "Stated outcome of this step",
    "judgment": "Match | Reasonable |
    Wrong | N/A"
  }
]
```

Your task is to reformat the following solution into discrete reasoning steps, and evaluate each step based on the ground truth.
Input:

```
[Problem]

{question}

[Solution]

{solution}

[Correct Answer]

{answer}

[Ground Truth Information]

{gt\_annotation}
```

C.2.3 Evaluation prompts for recall calculation

The prompt for recall calculation is:

Prompt for calculating recall for CoT outputs

You are an expert system for verifying solutions to medical image-based problems. Your task is to match the ground truth middle steps with the provided solution:

# Input Format:

1. Problem: The original question/task.
2. A Solution of a model.
3. Ground Truth: Essential steps required for a correct answer.

# Matching Process:
You need to match each ground truth middle step with the solution. Match Criteria:

- The middle step should match in the content or is directly entailed by a certain content in the solution.
- For subjective or descriptive steps such as image feature descriptions, treatment suggestions, disease causes, or cellular functions, match leniently: A step is considered "Matched" if the overall meaning largely overlaps with the solution and there is no contradiction, even if wording differs. Exact wording or structure is not required as long as the clinical implication is preserved.
- For objective steps such as specific diseases, lesion names, or image modalities, match more strictly: The terminology must refer to the same medical concept, though phrasing may differ (e.g., "retinal detachment" vs. "detached retina" is acceptable). Partial overlap is permitted, but the key meaning cannot be changed.

In all cases, evaluate whether each ground truth step is represented in the solution, either explicitly or with clear implication.
# Output Format:
JSON array of judgments:

```
[
  {
    "step_index": <integer>,
    "judgment": "Matched" | "Unmatched"
  }
]
```

# Additional Rules:

1. Only output the JSON array with no additional information.
2. Judge each ground truth middle step in order, without omitting any step.

Here is the problem, answer, solution, and the ground truth middle steps:

```
[Problem]

{question}

[Answer]

{answer}

[Solution]

{solution}

[Ground Truth Information]

{gt_annotation}
```

C.2.4 Evaluation prompts for step order recognition

When computing CoT consistency, it is necessary to determine the order of the reasoning steps in the model's output. This requires first classifying the type of each step. Our prompt is as follows:

Prompt for step order recognition

Our prompt consists of two main parts: a system instruction section (`system_prompt`) and an output format section (`OUTPUT_FORMAT`). The system part defines how the AI should analyze medical image responses, while the output format specifies the JSON structure, with the AI response to be analyzed being passed into the system through the `text` parameter:

```
{
  "modality_reasoning_segments":
  [AI's major segments that primarily focus on imaging
  techniques/examination methods],
  "observation_reasoning_segments":
  [AI's major segments that primarily focus on describing
  what is directly visible in the image],
  "conclusion_reasoning_segments":
  [AI's major segments that primarily focus on making
  definitive identifications, diagnoses,
  or final determinations],
  "knowledge_reasoning_segments":
  [AI's major segments that primarily focus on external
  clinical knowledge/context beyond
  what's visible],
  "modality_first_position":
  [character position where first modality-focused
  segment appears],
  "observation_first_position":
```

```
  [character position where first observation-focused
  segment appears],
  "conclusion_first_position":
  [character position where first conclusion-focused
  segment appears],
  "knowledge_first_position":
  [character position where first knowledge-focused
  segment appears],
  "modality_reasoning_order":
  [1-4 based on which type of segment appears first,
  0 if not present],
  "observation_reasoning_order":
  [1-4 based on which type of segment appears first,
  0 if not present],
  "conclusion_reasoning_order":
  [1-4 based on which type of segment appears first,
  0 if not present],
  "knowledge_reasoning_order":
  [1-4 based on which type of segment appears first,
  0 if not present],
  "total_segments":
  [total number of major reasoning segments AI used],
  "reasoning_pattern": "[A simple, high-level sequence
  of the primary reasoning categories, e.g.,
  ’Modality -> Observation -> Conclusion’]"
}
```

System Prompt: You are an expert AI reasoning analysis assistant. Analyze AI responses to medical image questions by identifying the AI’s own major logical segments and categorizing each segment by its PRIMARY focus.

# ANALYSIS INSTRUCTIONS:

1. Disregard any purely introductory or framing sentences (e.g., “I’ll analyze this image...”). Only analyze segments that contain substantive reasoning.
2. Respect the AI’s own major structural divisions (steps, sections, or natural paragraph breaks).
3. Categorize each major segment by its single, dominant reasoning type.
4. For the reasoning_pattern field, create a concise, high-level sequence of the primary categories. Do not include step numbers or repeat categories for consecutive segments of the same type.

SEGMENT CATEGORIES:

- MODALITY_REASONING: Segments about imaging techniques, examination methods, image types, technical aspects (e.g., “This is an endoscopic image”, “This appears to be a chest X-ray”)
- OBSERVATION_REASONING: Segments describing what is directly visible - anatomical structures, visual characteristics, findings without making definitive conclusions (e.g., “The tissue appears red”, “I can see circular structures”)
- CONCLUSION_REASONING: Segments making definitive identifications, diagnoses, final determinations, or conclusive statements about what something IS (e.g., “This is scoliosis”, “There is no bleeding present”)
- KNOWLEDGE_REASONING: Segments applying external clinical knowledge beyond the image - explaining what signs to look for, clinical context, background medical information (e.g., “Active bleeding would typically appear as...”, “Treatment options include...”)

# CRITICAL DISTINCTIONS:

- Simply mentioning “endoscopic image” within observation = MODALITY
- Describing visible red tissue = OBSERVATION
- Explaining what bleeding signs look like = KNOWLEDGE
- Stating “no bleeding present” = CONCLUSION

# Analyze the following: {text}

# CRITICAL: You must respond with ONLY valid JSON format. Do not include any other text before or after the JSON object.
Your output must be valid JSON in this exact format:
{OUTPUT_FORMAT}

D SUPPLEMENTARY TO EXPERIMENTS

D.1 SUPPLEMENTARY RESULTS

The radar plot for performances of some MLLMs on M3CoTBench is shown in Figure A7. Due to space limitations, we only reported the latency and efficiency metrics in the main text. Here, we present the average response time per question for each MLLM under both the direct and step-by-step settings, as shown in the Table below. As shown in Table A2, most MLLMs exhibit a significant increase in response time under the step-by-step (CoT) setting compared to the direct response setting. This is expected due to the inherently longer generation process of multi-turn reasoning. In general, closed-source commercial models tend to have higher latency than open-source models in both settings, likely because they employ larger architectures or more complex inference pipelines. For example, Gemini2.5-pro and GPT-4 variants demonstrate relatively high response times compared to smaller open-source models such as Qwen2.5-VL-7B and LLaVA-OV-7B. When comparing models of different scales, larger models usually incur higher latency due to increased computational cost; however, some exceptions exist, potentially due to optimization and deployment differences. Notably, the Llama-3.2-11B-Vision model shows an abnormally high latency in the direct setting, even exceeding that of its larger 90B counterpart, suggesting deployment inefficiencies rather than pure model complexity as the cause.

It is also important to note that public APIs are often affected by uncontrollable external factors such as server load, throttling policies, or background queuing. And the local experiments and API-based evaluations were conducted on different hardware environments, which may contribute to latency differences. Therefore, while the measurements reflect general trends in efficiency, they are subject to variability and may not precisely represent the models' inherent computational latency. This constitutes a limitation of our experiments.

Table A2: Comparison of the average response time per question for MLLMs under direct and step-by-step reasoning conditions. Optimal / sub-optimal results are highlighted in **bold** / underline.

Model	$T_{\textbf{direct}}$	$T_{\textbf{CoT}}$
LLaVA-OV-7B (Li et al., 2024)	0.7034	7.7822
LLaVA-CoT (Xu et al., 2024a)	5.5613	7.4875
Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	**0.5188**	8.0152
Qwen2.5-VL-72B-Instruct (Bai et al., 2025)	1.7144	13.3034
Llama-3.2-11B-Vision (Meta AI, 2024)	8.5518	9.6951
Llama-3.2-90B-Vision (Meta AI, 2024)	2.3694	12.5677
Gemini2.5-pro (Google DeepMind, 2024)	13.8923	23.3392
Claude-Sonnet-4 (Anthropic, 2024)	4.2489	11.4011
GPT-4o (OpenAI, 2024a)	2.3639	13.7521
GPT-4.1 (OpenAI, 2023)	2.2465	10.9657
LLaVA-Med (Li et al., 2023)	0.8703	**2.3244**
HuatuoGPT-Vision (Chen et al., 2024a)	<u>0.5310</u>	11.1364
HealthGPT (Lin et al., 2025)	0.6395	<u>5.3322</u>

D.2 CASE STUDY

D.2.1 EXAMPLE 1

Comparison of answers from Qwen2.5-VL-7B-Instruct and the annotated CoT steps.

Q: What is the most appropriate term to describe this finding? (Select one option)

A. Hyperplasia

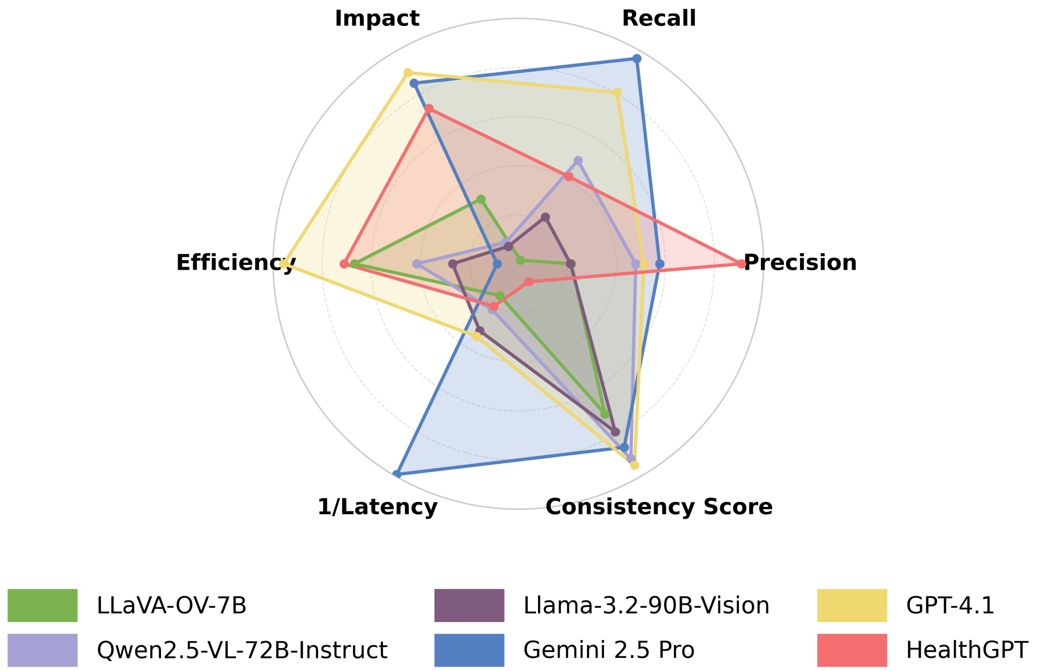


Figure A7: **Performances of some MLLMs on M3CoTBench:** Closed-source models such as Gemini 2.5 Pro perform well in terms of CoT correctness, negative CoT impact, consistency, and latency. However, Gemini 2.5's relatively low efficiency scores are likely due to the longer absolute inference time.

B. Dysplasia

C. Carcinoma in situ

D. Invasive carcinoma

Correct Answer: C.

The direct answer is C., and the CoT Answer is B.

The output of CoT prediction is shown below:

To determine the most appropriate term to describe the finding in the provided image, we need to analyze the characteristics of the tissue shown and compare them with the definitions of the given options.

Hyperplasia: This is an increase in the number of cells in a tissue without any abnormal changes in the cell structure or function. The image shows an increase in cellularity, but the cells appear to be normal in structure and function, which is not typical for hyperplasia.

Dysplasia: This is a precancerous condition where there is a change in the normal structure of cells, leading to an abnormal growth pattern. The cells in the image appear to have abnormal nuclei and cytoplasm, which is a characteristic of dysplasia.

Carcinoma in situ: This is a term used to describe a cancer that is confined to the epithelial layer and has not invaded the underlying stroma. The image shows a well-defined area with abnormal cells, but there is no evidence of invasion into the underlying stroma.

Invasive carcinoma: This is a cancer that has invaded the underlying stroma and has the potential to spread to other parts of the body. The image does not show any invasion into the underlying stroma, so this term is not applicable.

Given the analysis, the most appropriate term to describe the finding in the image is:

B. Dysplasia

The intermediate reasoning process is as follows:

The image shows an increase in cellularity.

The cells appear to have abnormal nuclei and cytoplasm, which is characteristic of dysplasia.

There is no evidence of invasion into the underlying stroma, so it is not carcinoma in situ or invasive carcinoma.

Hyperplasia typically does not involve abnormal cell structure or function, which is not the case here.

Final answer: B. Dysplasia

The annotated CoT steps are shown below:

1. Section stained with hematoxylin and eosin (H&E).

2. The cells show significant cytologic atypia and pleomorphism. The abnormal cells involve the entire thickness of the epithelium, but no invasion beyond the basement membrane. The architectural disturbance is pronounced, but the basement membrane remains intact.

3. The likely diagnosis is Carcinoma in situ.

Analysis:

When using CoT reasoning, Qwen7b gave an incorrect answer because it failed to integrate and weigh key medical diagnostic information accurately. Although it recognized abnormal nuclei and cytoplasm, it overlooked the critical criterion for carcinoma in situ that abnormal cells involve the entire thickness of the epithelium while the basement membrane remains intact. The CoT reasoning focused too much on cellular abnormalities and incorrectly classified the finding as dysplasia, without effectively distinguishing differences in the extent of cell involvement and structural integrity. Additionally, the model did not fully utilize the important information about the absence of basement membrane invasion, leading to a wrong conclusion. In contrast, the direct answer mode relied more on memorized mappings learned during training, which led to the correct answer. Based on the authoritative annotations you provided, explicitly emphasizing these key diagnostic features in the CoT prompt and guiding the model to verify these criteria during reasoning systematically could help build a more rigorous reasoning process and avoid such errors in complex medical questions.

D.2.2 Example 2

Comparison of answers from HuatuoGPT-Vision and the annotated CoT steps.

Q: Which type of cell is this most likely to be? (Select one option)

A. Basophil

B. Monocyte

C. Neutrophil

D. Eosinophil

Correct Answer: A.

The direct answer is A., and the CoT Answer is C.

The output of CoT prediction is shown below: *To determine which type of cell this is, let's analyze its characteristics step by step:*

Shape and Size

The cell in the image appears to be round with a large nucleus that occupies most of the cell volume.

Nucleus Characteristics:

Basophils: Have an oval or kidney-shaped nucleus.

Monocytes: Have a horseshoe- or kidney-shaped nucleus.

Neutrophils: Typically have multi-lobed nuclei (usually 3-5 lobes).

Eosinophils: Have bilobed or ring-like nuclei.

Cytoplasmic Granules:

Basophils: Contain dark purple granules due to basophilic substances like DNA/RNA and heparin.

Monocytes: May contain small granules but are generally less granulated than other cells.

Neutrophils: Have fine chromatin without distinct granules.

Eosinophils: Characterized by bright red-orange cytoplasmic granules containing eosinophil peroxidase enzyme.

Given these observations, it seems that the cell has a large rounded nucleus typical for neutrophils, yet lacks significant granulation seen typically in eosinophils. This suggests we might lean towards identifying it as either a monocyte or a neutrophil based on shape alone; however, further confirmation would require additional information such as staining properties specific to each cell type under microscopic examination.

Therefore, considering all aspects mentioned above, I conclude that the cell depicted here is most likely a neutrophil because of its characteristic multi-lobed nucleus despite lacking definitive granular features visible at this magnification level.

So my final answer is: C. Neutrophil

The annotated CoT steps are shown below:

1. Hematology or Cytology 2. The cell is roughly circular to oval in shape. The cytoplasm is densely filled with prominent, coarse, dark purple to violet-colored granules. These granules largely obscure the underlying nucleus, which appears as a dark, irregular central area. 3. The cell is a Basophil.

Analysis: In this case, the model's CoT reasoning led to an incorrect classification because it overlooked the most critical diagnostic feature emphasized in the annotated CoT: the presence of coarse, dark purple granules that densely fill the cytoplasm and partially obscure the nucleus, defining characteristics of a basophil. Instead, the model's reasoning assumed a clearly visible, multi-lobed nucleus and minimal granulation, which contradicts the image description in the annotated steps. This contrast highlights a key failure in the model's visual interpretation during CoT: while the direct answer correctly selected basophil, the CoT reasoning introduced assumptions that conflicted with observable features. The error demonstrates how CoT, when not properly grounded in domain-specific visual cues, can mislead the model away from an otherwise correct prediction.

E Limitation Discussion

E.1 Annotation Discrepancies Between Experts, AI, and Public Datasets

The question-answer pairs and CoT annotations were generated through collaboration between medical experts and AI, while also referencing labels from existing public datasets. In some cases, discrepancies arose between expert judgment and dataset labels. We generally prioritized the public dataset labels as the highest authority. However, we frequently encountered inconsistencies or potential errors in these labels. In such cases, we made efforts to verify through repeated reviews and multiple AI model assessments, but we cannot guarantee that every annotation step is fully accurate.

E.2 Disease-Specific Labels May Imply Unjustified Diagnostic Precision

Some annotations involve specific diseases (e.g., COVID-19, certain cancers), directly inherited from the original dataset labels. These labels may have been informed by additional contextual information unavailable in the image alone. In reality, making a definitive diagnosis from a single image is often not feasible, even for trained physicians. By retaining these disease-specific labels, the task may set an unrealistically high bar for MLLMs, possibly exceeding what is expected of human experts. To address this, we aimed to phrase our labels cautiously using formulations like "the most likely diagnosis is...".

E.3 Subjectivity in Expression May Affect Matching

Although we adopted relatively permissive matching criteria to account for variation in wording, certain annotation statements inevitably involve subjective interpretation, particularly when describing subtle visual findings or formulating likely diagnoses. These subjective elements can introduce variability in phrasing that, despite semantic similarity, may not be captured perfectly by automated matching methods. Furthermore, medical descriptions often allow for multiple valid expressions of the same observation, and differences in terminology, level of detail, or emphasis may lead to mismatches during evaluation. This issue is particularly relevant for open-ended reasoning tasks, where the boundary between correct and incorrect answers can be nuanced.

E.4 Evaluation Fully Conducted with GPT-4o

All evaluation of model outputs was conducted using GPT-4o. While GPT-4o has demonstrated strong performance in general reasoning and medical question answering, it remains an AI system with inherent limitations. In complex or ambiguous cases, the model may misinterpret medical terminology, overlook subtle differences between options, or apply inconsistent grading criteria. Additionally, its judgments may be influenced by prompt wording or prior context, leading to potential evaluation bias. The absence of human cross-validation means that certain edge cases could be mis-scored, especially in domains requiring precise domain knowledge, such as pathology or hematology.

For evaluation circularity concerns, although using a greater variety of models might lead to further improvement, the current annotation workflow is already effective in ensuring high-quality annotations while minimizing model bias. Specifically, by integrating two models, GPT-4o and Gemini-2.5-Pro, through multiple processing steps and incorporating manual expert correction, the risk of dominance by a single model has been significantly reduced. Experimental results also show that GPT-4o, which participated in the annotation, was not the top performer in the evaluation, which in turn serves as evidence that circular evaluation bias has been effectively controlled. Moreover, the final evaluation is based on comparing outputs with the annotated ground truth, rather than relying on the model to independently generate judgments, further reducing the risk of circularity.

E.5 No Inter-annotator Agreement Scores are Reported

Inter-annotator agreement scores: Because this workflow is not fully parallel, we acknowledge that inter-annotator agreement scores are not reported, which is a limitation of this study. However, the multi-stage review process, combining initial student review, multi-model automated checks, targeted expert verification, consensus discussions, and final read-through, ensures high-quality annotations while minimizing bias from any single reviewer or model. This careful workflow allows us to produce reliable reference reasoning chains suitable for evaluating MLLMs in medical image understanding.

E.6 No multiple experimental runs, and no confidence intervals were reported.

Due to cost and time constraints, this study only conducted a single evaluation and did not report confidence intervals or significance tests. We acknowledge that repeating experiments and reporting confidence intervals would provide more rigorous and reliable results. In future versions, we plan to include multiple runs and statistical significance analyses to further strengthen the robustness of our findings.

E.7 Limited Exploration of Prompts and Ablation Studies

In this study, we did not conduct a comprehensive exploration of alternative prompting strategies or perform extensive ablation experiments to evaluate the impact of prompt design choices systematically. Variations such as adjusting the level of detail, explicitly guiding reasoning steps, or introducing domain-specific constraints could potentially influence model performance. Similarly, ablation studies, such as removing specific reasoning cues, altering input formatting, or testing under different context lengths, might have provided more profound insights into model behavior. The absence of these experiments limits our ability to fully characterize how sensitive the results are to prompt engineering and task setup.

F Social Impact Discussion

The proposed M3CoTBench benchmark carries several important implications for the development and evaluation of medical AI systems:

F.1 Advancing interpretable medical AI

By explicitly evaluating the reasoning chains of MLLMs, M3CoTBench encourages transparency in how models arrive at their predictions. Understanding intermediate reasoning steps allows researchers and clinicians to better align AI behavior with clinical decision-making processes, fostering trust and supporting responsible deployment in medical research and practice. In high-stakes medical applications, interpretability is critical: clinicians can verify whether model reasoning is consistent with established diagnostic criteria, and researchers can identify failure modes that may not be apparent from final predictions alone.

F.2 Improving model evaluation in medical AI

Most existing benchmarks focus solely on final predictions, overlooking the reasoning process that leads to those outcomes. M3CoTBench fills this gap by providing a structured framework to assess the correctness, consistency, and efficiency of CoT reasoning across diverse medical imaging tasks. This enables a more nuanced analysis of model performance, highlighting specific strengths and weaknesses in reasoning patterns that are essential for complex diagnostic scenarios. By systematically evaluating intermediate steps, M3CoTBench supports the development of models that are not only accurate but also capable of robust and verifiable decision-making.

F.3 Promoting rigorous development of trustworthy AI systems

By emphasizing the evaluation of reasoning quality rather than only accuracy, the benchmark guides the design of models that are not only correct but also interpretable and reliable. This focus on transparent reasoning can help mitigate risks associated with opaque AI decisions in clinical settings, enabling more accountable AI deployment. Moreover, by providing standardized metrics for reasoning quality, M3CoTBench encourages best practices in medical AI development, fostering the creation of models that adhere to both technical and ethical standards.