InfiMM-Eval: Complex Open-ended Reasoning Evaluation for Multi-modal Large Language Models

Anonymous ACL submission

Abstract

Multi-modal Large Language Models are increasingly prominent due to their superior reasoning abilities to excel at complex tasks. Prevailing benchmarks related to multi-modal reasoning attempt to assess MLLMs through *yes/no* or multi-choice questions, which by design can introduce position bias and overlook the intermediate reasoning process, thereby rendering the results less convincing. To this end, we systematically categorize the reasoning tasks into deductive, abductive and analogical reasoning, and introduce InfiMM-Eval, a manually curate benchmark featuring 279 diverse and nuanced reasoning questions across these categories. The questions are designed to be fully *open-ended* to better represent the characteristics of generative models. To mitigate the challenge of answering complex reasoning questions, we encourage models to generate intermediate reasoning steps. These steps are incorporated into the evaluation protocol to reduce bias towards plausible guesses or responses that lack definitive answers, while facilitating the assessment of more nuanced reasoning skills. This evaluation scheme closely resembles the method by which humans evaluate exams in real-world settings, enabling a more reliable assessment. We evaluate a large selection of trending MLLMs to reveal the discrepancies in reasoning abilities between open-source and proprietary MLLMs. Additionally, we conduct a comprehensive analysis of three reasoning related factors, highlighting potential directions for further research in elevating MLLMs in reasoning tasks.

1 Introduction

Exhibiting exceptional proficiency in a wide range of NLP tasks (Devlin et al., 2018; Radford et al., 2019), large language models (LLMs) have led to the development of multi-modal large language models (MLLMs), which incorporate multi-modal perception, primarily visual information, into language models for more versatile content understanding and generation across domains (Alayrac et al., 2022; Rombach et al., 2022; Driess et al., 2023; Ghosal et al., 2023). Leading proprietary models such as Palm-e (Driess et al., 2023), Flamingo (Alayrac et al., 2022), RT-2 (Brohan et al., 2023), and GPT-4V(ision) (OpenAI, 2023b) have exemplified the extensive applicability and promising potential of MLLMs. The open-source community has also contributed significantly to the field through the development of innovative architectures and the creation of curated instruction fine-tunning datasets, including MiniGPT-4 (Zhu et al., 2023a), LLaVA (Liu et al., 2023b), IDEFICS (Laurençon et al., 2023), etc. Each model provides distinct insights, exploring a variety of data recipes and approaches on multi-modal alignment.

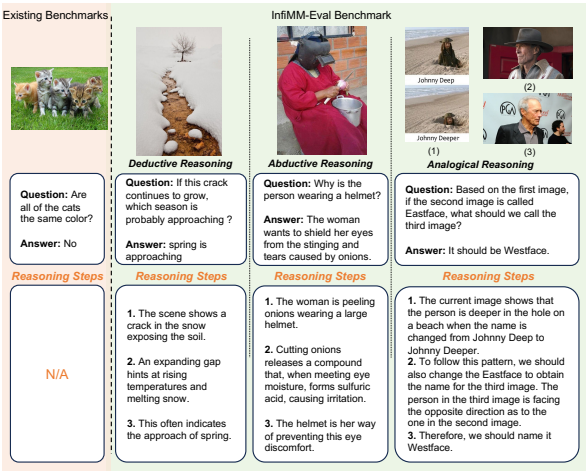

Figure 1: Comparison between existing MLLM benchmarks and InfiMM-Eval. **Left**: Existing benchmarks involve basic reasoning tasks with simple responses. **Right**: InfiMM-Eval consists of deductive, abductive, and analogical reasoning, each of which includes one or multiple images, one question and one answer with nuanced intermediate reasoning steps.

Reasoning is a key factor for human-level intelligence especially in complex tasks (McCarthy, 2007; Darwiche, 2018), yet it is challenging to

evaluate and often escalates unpredictably, requiring specialized benchmarks such as ARB (Sawada et al., 2023), ARC (Clark et al., 2018), and GSM8k (Cobbe et al., 2021). The desire for specialized reasoning benchmarks for MLLMs is even more critical considering the complexity of multimodal perception (Zellers et al., 2019a).

Recent advancements in the MLLMs field have led to the establishment of comprehensive evaluation benchmarks such as MME (Fu et al., 2023), MMBench (Liu et al., 2023c), SeedBench (Li et al., 2023b), and MathVista (Lu et al., 2023). While reasoning ability is an important factor assessed in these benchmarks, there lacks a consistent categorization of reasoning capabilities which is critical for generating fine-gained analysis and comprehensive insights. Existing benchmarks may not fully challenge the limits of advanced models like GPT-4V due to their reliance on simple responses or multiple-choice formats, which do not adequately reflect the complexity and format diversity of reasoning tasks. Additionally, such constrained formats coupled with the lack of intermediate reasoning steps render the results susceptible to plausible short answers and cases when no definite answers are generated. This highlights the need for a rigorous and holistic benchmark to accurately assess the reasoning capabilities of advanced MLLMs.

To address the issues identified above, we introduce the **InfiMM-Eval** benchmark which is designed to evaluate open-ended complex multimodal reasoning problems. Drawing on the work of (Conner et al., 2014) in the field of logical reasoning, we categorize samples into three reasoning paradigms: deductive, abductive, and analogical reasoning. The example of each category is shown in Figure 1. This categorization encompasses a broad range of practical applications in reasoning, and thus offers comprehensive insights into the reasoning capabilities of MLLMs. In addition to only offering question-answer pairs like other benchmarks, InfiMM-Eval incorporates explicit reasoning steps that delineate the derivation of ground truth answers. This approach not only minimizes the potential for results to be swayed by fortuitous guesses but also embraces open-ended responses, which are inherently more aligned with the complexities encountered in real-world situations. This enhancement ensures a more precise and practical evaluation, especially in scenarios that demand intricate reasoning. To the best of our knowledge, InfiMM-Eval represents the first open-ended multimodal QA benchmark featuring manually curated intermediate reasoning steps as ground truth.

The inclusion of reasoning steps facilitates the creation of a more sophisticated evaluation protocol. Our evaluation protocol is designed following the rubric grading format that is widely used in exams, where the response receives full marks for a directly correct answer, or partial scores based on the relevance and logic of its intermediate reasoning steps. This method not only underscores the model's proficiency in generating correct answers, but also provides a thorough analysis of its decision-making process for a fully acurate evaluation. We employ an LLM-based evaluator to execute this evaluation protocol for better efficiency. With a collection of 279 high-quality and diverse samples across three reasoning categories, it is our hope that this benchmark will serve a cornerstone in the MLLMs' reasoning evaluation, similar to HumanEval (Chen et al., 2021) in code generation. Our contributions can be summarized as follows:

- We present InfiMM-Eval, a manually curated high-quality benchmark, featuring complex reasoning questions tailored to fully assess the MLLMs.
- We propose a robust protocol to evaluate open-ended model responses. By integrating intermediate reasoning steps with final answers, the evaluation results are more accurate and more aligned with real-world scenarios.
- We evaluate a broad spectrum of leading MLLMs on InfiMM-Eval, and analyze the related factors to the reasoning capabilities through extensive ablation studies.

2 Related Work

2.1 Multi-modal LLMs

The evolution of LLMs has inspired research on integrating visual signal into LLMs. For example, Flamingo (Alayrac et al., 2022) integrates the Perceiver Resampler (Jaegle et al., 2021) and gated attention modules onto LLMs, bridging visual encoders and LLMs, thereby proving highly effective in in-context learning capability for vision-language tasks. Other large-scale models like Palm-e (Driess et al., 2023), RT-2 (Brohan et al., 2023), and GPT-4V(ision) (OpenAI, 2023b) have also underscored the extensive applicability and promising potential of MLLMs.

Various smaller-sized MLLMs have emerged recently. Mini-GPT4 (Zhu et al., 2023b) utilizes the instruction-tuned Vicuna (Chiang et al., 2023), and fine-tunes a linear layer to align vision and language representations. LLaMA-Adapter (Zhang et al., 2023b) introduces a lightweight adapter to enable the adaptability of LLaMA to visual inputs. BLIP-2 (Li et al., 2023d) incorporates the Q-Former, adding a crucial alignment stage to connect the frozen LLM with the visual modality, notably excelling in Visual Question Answering (VQA) tasks. InstructBLIP (Dai et al., 2023) focuses on fine-tuning the Q-Former using diverse instruction tuning datasets, enhancing its performance in visual scene comprehension and visual dialogues. In contrast, Otter (Li et al., 2023a), refines the OpenFlamingo (Awadalla et al., 2023) for improved instruction-following capabilities and more effective usage of in-context samples. Multimodal-CoT (Zhang et al., 2023c) integrates chain-of-thought (Kojima et al., 2022; Wei et al., 2022b) into the multimodal domain, showcasing robust results on the ScienceQA benchmark. MMICL (Zhao et al., 2023b) tackles the challenges posed by multimodal inputs with multiple images, targeting intricate multi-modal prompts and detailed text-to-image references. LLaVA (Liu et al., 2023b) employs a simple linear connector and fine-tunes the entire LLM to boost performance. The upgraded version, LLaVA-1.5 (Liu et al., 2023a), incorporates large-scale instruction tuning and high-resolution images, resulting in superior performance across multiple benchmarks.

2.2 MLLM Evaluation Benchmarks

Different vision-language benchmarks have been introduced to evaluate the specific reasoning capabilities of MLLMs. For instance, Winoground (Thrush et al., 2022) assesses the visual-linguistic compositional reasoning, RAVEN (Zhang et al., 2019) focuses on relational and analogical reasoning, OK-VQA (Marino et al., 2019) examines reasoning with external knowledge, and VCR (Zellers et al., 2019b) evaluates visual commonsense reasoning related to people in video frames. Other benchmarks, such as TextVQA (Singh et al., 2019), FigureQA (Kahou et al., 2018), and ScienceQA (Saikh et al., 2022), have also made significant contributions by addressing reasoning within diverse contexts. MathVista (Lu et al., 2023) provides a consolidated assessment of mathematical reasoning capabilities.

In addition to the above-mentioned reasoning-specific benchmarks, comprehensive benchmarks have been proposed, which also include assessments of various reasoning capabilities. For instance, MME (Fu et al., 2023) evaluates reasoning capabilities of commonsense reasoning, numeric calculation, text translation, and code understanding. MMBench (Liu et al., 2023c) assesses logical, attribute, and relation reasoning, while SEED-Bench (Li et al., 2023c) contains visual reasoning, action prediction, and procedure understanding. All above benchmarks use multiple-choice question format to simplify the evaluation process. As studied in (Zong et al., 2023), multiple-choice questions may include bias and additional hints, popular MLLMs are vulnerable to adversarial permutation in answer sets for multiple-choice prompting. On the other hand, scoring by final answer correctness only underestimates the importance of reasoning process, which is not enough to understand the models' reasoning capability.

Thus, open-ended benchmarks are needed to better align with the generative nature of recent MLLMs. However, traditional metrics, like CIDEr (Vedantam et al., 2015), SPICE (Anderson et al., 2016), etc. are not suitable for open-ended QA evaluation. Human evaluations are prohibitively costly. Luckily, (Chiang and Lee, 2023) suggest LLMs can be an alternative to human evaluators. Recent open-ended QA benchmarks for MLLMs, such as TouchStone (Bai et al., 2023c), VisIT-Bench (Bitton et al., 2023), and MM-Vet (Yu et al., 2023b), also employ LLM-based evaluators. This further demonstrates the reliability of LLM-based evaluators in such context.

2.3 Reasoning in MLLMs

Human reasoning, essential for intelligence, involves analyzing information to derive logical insights (Yu et al., 2023a; Huang and Chang, 2022; Walton, 1990). LLMs have demonstrated substantial reasoning abilities in NLP tasks, as evidenced in recent studies (Kojima et al., 2022; Huang and Chang, 2022; Wei et al., 2022a; Yao et al., 2022; Webb et al., 2023). Similar capabilities are observed in (Driess et al., 2023; OpenAI, 2023b). However, MLLMs research field lacks a systematic and unified framework for categorizing reasoning capability. Current benchmarks fragment reasoning into numerous task-specific categories, e.g. commonsense reasoning, math reasoning, code understanding, procedure understanding. Such cat-

egorization may potentially obscure a holistic understanding of the reasoning capacities of MLLMs. Our study advocates for a directional classification of reasoning in MLLMs, anchored in established logical principles (Bronkhorst et al., 2020; Dowden, 2018), focusing on deductive, abductive, and analogical reasoning, essential in human cognition. Detailed categorization and corresponding examples can be found in Appendix A.

In this work, we present InfiMM-Eval, an open-ended VQA benchmark specifically created to evaluate the reasoning abilities of MLLMs. This benchmark features systematic design and categorization of reasoning questions, aimed at comprehensively assessing MLLMs' reasoning capabilities.

3 InfiMM-Eval Benchmark

3.1 Data Collection

Compared with the extensive, automatically collected MLLM reasoning datasets as discussed in prior studies (Li et al., 2023a; Liu et al., 2023b; Zhao et al., 2023a), our InfiMM-Eval initiative is dedicated to the manual creation of a high-quality evaluation benchmark. This benchmark is particularly designed to evaluate the multi-step reasoning abilities increasingly evident in contemporary MLLMs. It specifically emphasizes deductive, abductive, and analogical reasoning, which are fundamental to routine human cognitive processes.

In alignment with this principle, the process of collecting data for our evaluation benchmark can be broadly categorized into the following steps:

Question and Answer Collection. Our methodology involved engaging eight annotators with advanced education level, each tasked with sourcing a wide range of images from varied scenarios. These images were sourced from a variety of platforms, including online platforms and existing public dataset, notably adopting 25 samples from MM-Vet (Yu et al., 2023b). The primary objective for these annotators was to create a comprehensive set of questions and answers. It was imperative that these questions were crafted to rigorously test the multi-step logical reasoning capabilities of MLLMs. To ensure the complexity of the task, the questions were designed to be intricate enough to preclude the possibility of immediate answers based purely on visual observation.

To ensure the robustness of this study, specific guidelines were established for the formulation of questions. Although the answers format were permitted a degree of openness, the questions themselves were required to have a single logic path. This means that despite the potential openness in responses, the line of reasoning to arrive at these answers should be fairly consistent among different individuals. For example, overly subjective questions like "What is your feeling when you see this image?" were excluded. These types of questions do not align with the standard of robustly eliciting a logical reasoning pathway.

Additionally, we instructed annotators to categorize each question from the following aspects. Each question is reviewed by a minimum of 3 annotators. The final categorization is determined through a majority vote to ensure the reliability.

- **Reasoning category:** In alignment with principles of logical reasoning, questions are to be classified into one of three categories: deductive, abductive, or analogical reasoning.
- **Question complexity:** The complexity of a question is assessed based on multiple criteria, including the number of logical steps required for resolution, the extent of knowledge needed, and the presence of any elements that might introduce confusion or misinterpretation. Our guidelines delineate questions into "High" and "Moderate" complexity levels, primarily based on the number of intermediate reasoning steps involved. Nevertheless, annotators are afforded discretion to apply their judgment in borderline cases.
- **Question intuitivity:** This dimension evaluates how intuitively one can grasp the essence of the question and the possible answers. Annotators have the liberty to classify questions as either "Intuitive" or "Counter-Intuitive" depending on their immediate perception of the question's clarity and the straightforwardness of its potential answers.

Quality Control. To guarantee the exceptional quality of our benchmark, we implemented a thorough cross-validation protocol. Each sample underwent validation by two independent annotators. Their evaluation is based on a comprehensive set of standards, which includes:

- **Appropriateness:** Each image and question is examined for inappropriate or offensive content, ensuring fairness, diversity, and suitability for a diverse audience.

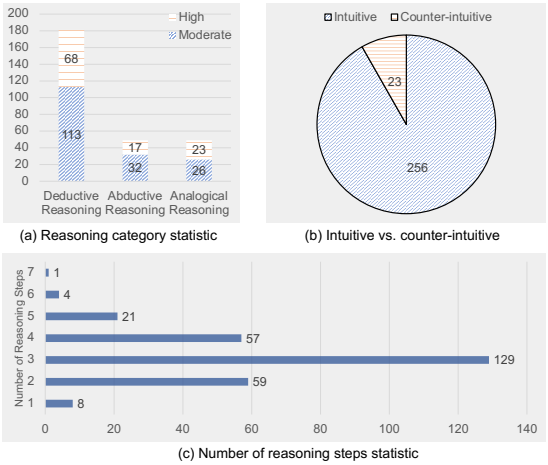

Figure 2: InfiMM-Eval benchmark statistics: (a) indicates distribution of reasoning categories and their respective reasoning complexity; (b) represents the statistic of counter-intuitive versus intuitive reasoning questions; and (c) shows the breakdown of the number of reasoning steps per question.

- **Consistency analysis:** The relationship between the question, answer, and reasoning steps are carefully evaluated to ensure they are logically aligned and coherent.
- **Image relevance:** This criterion assesses whether the image is essential for answering the question, thereby filtering samples where questions could be answered without the visual aid.
- **Complexity requirement:** Questions deemed overly simplistic, answerable by a cursory glance at the image without substantive logical engagement, were excluded.
- **Subjectivity and discrepancy:** If a question is found to be too subjective, or if the validators' answers significantly differ from the original answer, the question is either revised or removed.
- **Question format diversity:** We ensure a diverse representation of question formats, avoiding the overuse of any particular format of questions.

After rigorously applying these quality control measures in several review cycles, our benchmark was refined to include 279 high-quality samples. All samples satisfy our stringent criteria for accuracy, relevance, and cognitive challenge, ensuring a robust and reliable dataset.

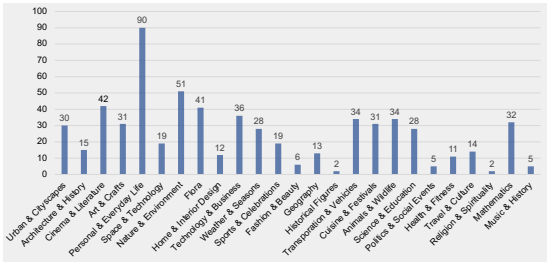

Figure 3: The distribution of visual content categories in InfiMM-Eval. A single image can encompass multiple visual content categories.

3.2 Dataset Statistics

In summary, our InfiMM-Eval benchmark consists of 279 manually curated reasoning questions, associated with a total of 342 images. Out of these, 25 images are adopted from MM-Vet, enriching the diversity and scope of the dataset.

We present a comprehensive statistical analysis of the dataset. Figure 2 (a) illustrates the distribution across various reasoning types: 49 questions pertain to abductive reasoning, 181 require deductive reasoning, and 49 involve analogical reasoning. Furthermore, the dataset is divided into two folds based on reasoning complexity, with 108 classified as "High" reasoning complexity and 171 as "Moderate" reasoning complexity. For both abductive and deductive reasoning categories, the ratio of "High" to "Moderate" questions reasoning complexity is approximately $1:2$, whereas for analogical reasoning, this ratio is closer to $1:1$. This distribution underscores the high quality of our benchmark. Notably, the dataset includes 23 questions that entail counter-intuitive reasoning (See Appendix for more details), further exemplifying the diversity of our benchmark, as depicted in Figure 2 (b). Additionally, as Figure 2 (c) indicates, about 76% (212 out of 279) of the reasoning questions require three or more steps to solve.

Figure 3 demonstrates the diversity of visual content in our image collection, categorized by GPT-4V into a predefined set of concepts.

3.3 Dataset Comparison

We provide a detailed comparison with other MLLMs reasoning benchmarks in Appendix C focusing on the aspects including data domain, data collection, answer format and whether intermediate reasoning steps are provided and considered. In summary, unlike other benchmarks, InfiMM-Eval features compiling questions from open-domain

real-world scenarios that involve more complex and unique logical reasoning processes. Additionally, our benchmark considers the accuracy of intermediate reasoning steps in the computation of the final metric. InfiMM-Eval is designed to complement existing benchmarks by offering an additional measure for evaluating the reasoning capabilities of MLLMs.

4 Experiments

In this section, we delineate the experimental settings to assess the reasoning capabilities in contemporary MLLMs. Specifically, we furnish a comprehensive description of evaluation baselines and protocols in section 4.1. Subsequent to this, we conduct thorough evaluations and ablation studies on a range of MLLMs using our InfiMM-Eval dataset, as detailed in section 4.2. The prompts we used for evaluating each model can be found in Appendix E.

4.1 Evaluation Protocol

Considering the open-ended nature of question-answering in the InfiMM-Eval benchmark and the generative capabilities of modern MLLMs, it becomes clear that solely assessing answer correctness is insufficient, e.g. in Figure 4. In line with recent studies (Bai et al., 2023c; Bitton et al., 2023; Yu et al., 2023b), we also employ LLMs as evaluators. However, our approach is distinct in its integration of both questions and answers, as well as the ground-truth and model-predicted reasoning steps into the LLM prompt. The inclusion of structured reasoning steps into the LLM context facilitates the accommodation of diverse model outputs and establishes a comprehensive and justified scoring system. As elaborated in section 1, our grading protocol awards full marks for direct correctness, with partial scores assigned based on the relevance and logic of reasoning steps. This method evaluates not only the model's accuracy in answer generation but also offers a an in depth analysis of its decision-making process, illuminating its reasoning pathways. For any given question q, its score s_q falls within the range of $[0, 1]$. The overall score S over the entire dataset, which includes considerations of reasoning complexity detailed in section 3.2, is calculated as

$$S = \frac{\sum_{x \in M} s_x + 2 \cdot \sum_{y \in H} s_y}{|M| + 2 \cdot |H|} \times 100\%, \quad (1)$$

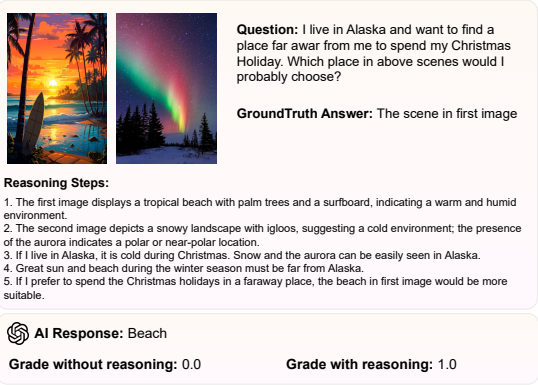

Figure 4: In this example, model can successfully answer the question, however, due to the nature of open-ended response, the model's response cannot be judged correctly solely based on question and answer.

where M and H denote the sets of questions categorized as having "Moderate" and "High" reasoning complexity, s_x and s_y denote score of each question belong to "Moderate" or "High" categories respectively, a coefficient of 2 is applied to "High" complexity category to balance the number of samples of each complexity category.

4.2 Benchmarking MLLMs on InfiMM-Eval

We evaluate a diverse range of MLLMs on InfiMM-Eval for their reasoning abilities, including GPT-4V (OpenAI, 2023a), LLaVA-1.5 (Liu et al., 2023b), Otter (Li et al., 2023a), MiniGPT-v2 (Zhu et al., 2023a), InstructBlip (Dai et al., 2023), Blip-2 (Li et al., 2023d), LLaMA-Adapter-V2 (Zhang et al., 2023b), InternLM-XComposer (Zhang et al., 2023a), QWen-VL-Chat (Bai et al., 2023a), and Fuyu (Bavishi et al., 2023).

The principal findings are encapsulated in Table 1, derived from employing the most effective prompt strategy for each model. Among all evaluated MLLMs, GPT-4V is particularly noteworthy, exhibiting unparalleled proficiency across all reasoning domains and complexities, with an overall reasoning score of 77.44. In the realm of open-source MLLMs, InfiMM-v1 is distinguished as the front-runner with the highest 41.32 overall score, marginally surpassing SPHINX-v2. Additionally, we observe that models fine-tuned with explicit instructions, display superior performance compared to their solely pretrained counterparts, exemplified by models such as Otter and OpenFlamingo-v2.

Table 1 further provides a granular breakdown of scores, reflecting the varied reasoning capabilities of the MLLMs. GPT-4V continues to exhibit

Table 1: Results for various MLLMs. Open-source models best performances are indicated with underlines.

MLLMs	LLM	IFT	Reasoning Category			Reasoning Complexity		Overall
			Deductive	Abductive	Analogical	Moderate	High	
OpenFlamingo-v2 (Awadalla et al., 2023)	MPT-7B (Team, 2023b)	No	8.88	5.3	1.11	9.47	4.72	6.82
MiniGPT-v2 (Zhu et al., 2023a)	LLaMA2-7B (Touvron et al., 2023)	Yes	11.02	13.28	5.69	14.45	7.27	10.43
Fuyu-8B (Bavishi et al., 2023)	Persimmon-8B (Elsen et al., 2023)	No	16.42	21.49	7.78	23.06	9.91	15.7
BLIP-2 (Li et al., 2023d)	OPT-2.7B (Zhang et al., 2022)	No	22.76	18.96	7.5	24.05	14.18	19.31
InternLM-XComposer-VL (Zhang et al., 2023a)	InternLM-7B (Team, 2023a)	Yes	26.77	35.97	18.61	39.13	17.18	26.84
InstructBLIP (Chung et al., 2022)	FLAN-T5-XXL (Chung et al., 2022)	Yes	27.56	37.76	20.56	40.64	18.09	28.02
LLaMA-Adapter V2 (Gao et al., 2023)	LLaMA-7B (Touvron et al., 2023)	No	28.7	46.12	22.08	41.33	21.91	30.46
Otter (Li et al., 2023a)	LLaMA-7B	Yes	22.49	33.64	13.33	35.79	12.31	22.69
mPLUG-Owl2 (Ye et al., 2023)	LLaMA-7B	Yes	23.43	20.6	7.64	28.79	13.18	20.05
IDEFICS-9B-instruct (Laurençon et al., 2023)	LLaMA-7B	Yes	22.99	34.63	20.56	34.45	16.73	24.53
Emu (Sun et al., 2023)	LLaMA-13B	Yes	28.9	36.57	18.19	36.18	22.0	28.24
LLaVA-1.5 (Liu et al., 2023b)	Vicuna-13B (Chiang et al., 2023)	Yes	30.94	47.91	24.31	47.4	21.0	32.62
CogVLM-Chat (Wang et al., 2023)	Vicuna-7B	Yes	36.75	47.88	28.75	55.67	22.5	37.16
Qwen-VL-Chat (Bai et al., 2023a)	Qwen-14B (Bai et al., 2023b)	Yes	37.55	44.39	30.42	46.61	**30.09**	37.39
SPHINX-v2 (Lin et al., 2023)	LLaMA2-13B	Yes	**42.17**	**49.85**	20.69	54.85	27.31	39.48
InfiMM-v1 (Team, 2024)	LLaMA2-13B	Yes	41.69	49.70	**32.36**	**61.81**	25.09	**41.32**
GPT-4V (OpenAI, 2023a)	GPT-4	Yes	**74.86**	**77.88**	**69.86**	**93.98**	**58.98**	**74.44**

its dominance across all reasoning dimensions. Interestingly, most open-source models lag behind GPT-4V, especially in analogical reasoning, which requires not only the detailed comprehension of image content, but also the ability to transfer knowledge from known instances to analogous situations.

To dive deeper, we stratify questions into two levels of complexity: "Moderate" and "High". See Appendix B for visualization of examples with varied reasoning complexities. It is noteworthy that GPT-4V consistently outperforms in addressing both moderate and high-complexity questions. Among the open-source models, InfiMM-v1 notably excels in managing moderate complexity questions, whereas Qwen-VL-Chat is particularly adept at handling high-complexity questions.

4.3 Factors Related to Reasoning Ability

Because InfiMM-Eval benchmark provides an accurate evaluation on MLLMs' reasoning ability, we further conduct a more fine-grained analysis on examining the impact of different MLLM techniques and factors over reasoning ability, including Chain-of-Thought (CoT) (Kojima et al., 2022; Wei et al., 2022b), in-context learning (Li et al., 2023a; Alayrac et al., 2022; Dong et al., 2022), and different model scales.

4.3.1 Results with Chain-of-Thought Prompt

In this section, we present a quantitative analysis examining the impact of CoT prompting on MLLMs. The results are detailed in Table 2.

We adopt a CoT prompting technique similar to that described in (Kojima et al., 2022) by appending "Let's think step by step" to the end of each question to enhance the reasoning capabilities of the model. Our results indicate varied performance changes across different models. Open-source models generally exhibit a minimal differences in performance, whereas GPT-4V exhibits a notable improvement of 3.7 with CoT prompts. We hypothesize that this phenomenon is attributed to differences in language model size and data quality during the instruction-finetuning (IFT) stage of model training. The majority of open-source MLLMs are limited by smaller language models, typically with less than 14 billion parameters, inherently constraining their reasoning abilities. Additionally, the scale and quality of the IFT datasets, commonly used in open-source MLLMs, influence the outcome significantly. A considerable portion of the IFT data, primarily sourced from VQA (Goyal et al., 2017), lacks in reasoning and commonsense knowledge.

Table 2: Comparative evaluation results of MLLMs with and without Chain-of-Thought prompts.

MLLMs	CoT	Deductive	Abductive	Analogical	Overall
BLIP-2	w/o	22.13	18.66	5.69	18.52
	w	22.76	18.96	7.5	19.31
InstructBLIP	w/o	25.2	34.48	16.94	25.27
	w	27.56	37.76	20.56	28.02
LLaVA-1.5	w/o	30.94	47.91	24.31	32.62
	w	31.18	48.51	22.78	32.6
Qwen-VL-Chat	w/o	38.55	45.91	22.5	36.82
	w	37.55	44.39	30.42	37.39
GPT-4V	w/o	69.88	77.88	67.08	70.72
	w	74.86	77.88	69.86	74.44

Table 3: Results with in-context learning example.

MLLMs	ICL	Deductive	Abductive	Analogical	Overall
Otter	w/o	22.49	33.64	13.33	22.69
	w	23.25	32.58	14.31	23.18
Qwen-VL-Chat 7B	w/o	33.73	46.82	30.28	35.32
	w	38.84	44.39	27.22	37.62
GPT-4V	w/o	74.86	77.88	69.86	74.44
	w	74.82	80.45	64.17	73.8

This raises a question about the feasibility of replicating of CoT's success in multimodal contexts.

4.3.2 Results with In-Context Learning

To examine the impact of in-context learning on the reasoning abilities of MLLMs, we selected three models ranging from the high-performing GPT-4V, alongside leading open-source models such as QWen-VL-Chat and Otter. It is noteworthy that only Otter incorporates in-context learning during its training phase. We randomly select an example from our dataset and concatenate it to the prompts during inference for each query, so that the selected example can help refine the reasoning process and ideally enhance the performance of these models.

As shown in Table 3, it is notable that the integration of in-context learning technique does not enhance, and may slightly impair, the performance of the GPT-4V. In contrast, marginal improvements in performance are observed in the Otter and Qwen-VL-Chat. These results underscore the complex and diverse nature of the benchmark employed in this study. Specifically, for the high-performing GPT-4V, the randomly selected ICL examples might significantly diverge from the test samples. Conversely, for models with smaller language encoders, such as Otter and Qwen-VL-Chat, which initially demonstrate inferior performance compared to GPT-4V, the inclusion of ICL examples potentially aids in the reasoning process, albeit the impact is relatively limited.

4.3.3 Results with LLMs of Varied Scales

Table 4 presents the evaluation results of MLLMs employing LLMs of different scales. The size of the LLMs is a critical determinant in augmenting the reasoning capabilities of MLLMs. For instance, considering Qwen-VL(Bai et al., 2023b) as a case study, there is a noticeable increase in the overall reasoning score concurrent with the expansion of the LLM's size. Specifically, when the model's size is increased from 7B to 14B parameters, its reasoning score increases from 35.32 to 37.39.

Furthermore, we also report the reasoning capability of standalone language models, such as Vicuna (Chiang et al., 2023) and GPT4 (OpenAI, 2023b), by replacing images with their corresponding textual descriptions. Prompting GPT-4 directly with only the question resulted in a reasoning score close to 0, as shown in the first row of Table 4). This suggests that the inclusion of visual elements is essential for accurate and effective responses. As we increase the model size of the LLaMA, from 7B to 70B, there is a noticeable improvement in reasoning scores when utilizing high-quality image descriptions generated by GPT-4V. The application of CoT markedly enhances the performance of SOLAR-0-70B, elevating its scores from 48.71 to 55.59. In contrast, this technique does not produce proportionate enhancements in smaller models, such as those with 7B and 13B.

Table 4: Results of MLLMs with varied LLM sizes.

Models	LLM	Caption	Deductive	Abductive	Analogical	Overall
GPT-4	GPT-4	-	5.82	5.0	2.5	5.06
Vicuna-7B	LLaMA-7B	GPT-4V cap.	38.01	48.98	30.0	38.53
Vicuna-13B	LLaMA-13B	GPT-4V cap.	34.42	58.78	34.69	38.75
SOLAR-0-70b	LLaMA-70B	GPT-4V cap.	48.56	64.49	33.47	48.71
GPT-4	GPT-4	GPT-4V cap.	54.59	66.73	45.1	55.05
Vicuna-7B(CoT)	LLaMA-7B	GPT-4V cap.	34.42	58.78	34.69	38.75
Vicuna-13B(CoT)	LLaMA-13B	GPT-4V cap.	39.39	46.33	34.08	39.68
SOLAR-0-70B(CoT)	LLaMA-70B	GPT-4V cap.	54.7	67.14	47.35	55.59
GPT-4(CoT)	GPT-4	LLaVA1.5 cap.	23.29	44.7	29.17	29.74
GPT-4(CoT)	GPT-4	GPT-4V cap.	55.75	66.53	51.22	56.85
LLaVa-1.5	LLaMA2-7B-Chat	-	27.8	33.28	21.11	27.51
	LLaMA2-13B-Chat	-	30.94	47.91	24.31	32.62
Qwen-VL-Chat	Qwen-7B	-	33.73	46.82	30.28	35.32
	Qwen-14B	-	37.55	44.39	30.42	37.39

The GPT-4 model demonstrates optimal reasoning performance when it employs the CoT technique in conjunction with image descriptions generated by GPT-4V. A significant reduction in performance is noted when these descriptions are substituted with those produced by LLaVA-1.5. Further analysis reveals that the detailed information in GPT-4V's descriptions, including OCR and extensive commonsense knowledge, is crucial for enhancing the "*multi-modal*" reasoning capabilities of standalone LLMs.

For more ablation studies, please see Appendix F.

5 Conclusion

In this paper, we introduce InfiMM-Eval, a comprehensive benchmark specifically designed to evaluate complex reasoning capabilities in MLLMs. InfiMM-Eval incorporates questions and answers for each data sample as well as detailed reasoning steps. We employ GPT-4 for the assessment and grading. Our evaluation covers a broad spectrum of MLLMs. We conduct extensive ablation studies to discern performance disparities among these models. The findings reveal that GPT-4V attains an overall score of 74.44. It is noteworthy that the top-performing open-source MLLMs still largely fall behind GPT-4V. InfiMM-Eval is poised to be a foundational benchmark for future enhancements in advancing reasoning capabilities of MLLMs.

6 Limitations

In this section, we delve into the possible constraints and shortcomings of the current InfiMM-Eval benchmark. Furthermore, we identify and suggest potential pathways for enhancement.

- **Expanding reasoning categories:** The InfiMM-Eval benchmark represents an initial endeavor to scrutinize the capability of deductive, abductive, and analogical reasoning in contemporary MLLMs. Notwithstanding, the spectrum of human reasoning transcends these categories, incorporating more complex forms such as inductive and causal reasoning. Future iterations of this benchmark aim to encompass a broader range of reasoning categories, thereby facilitating a more comprehensive assessment of reasoning capabilities.

- **Enhancing evaluation experiences:** Due to the size of the benchmark and the nature of LLM-based evaluation protocol, we have decide to only release images and corresponding questions, while maintaining an evaluation server that allows the public to submit model predictions to obtain final scores. This approach ensures that intermediate steps and answers remain confidential to prevent data leakage. We will conduct further research to develop a more refined metric and evaluation protocol, aims to provide intermediate reasoning step scores to better diagnostic MLLMs without compromising data.

7 Ethical considerations

This work proposes an MLLMs evaluation benchmark, with potential risks of being misused for assessing models trained for harmful usage, e.g. malicious web agent. Our benchmark design and methodology aim to minimize these risks, ensuring their impact remains low.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023a. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. 2023c. Touchstone: Evaluating vision-language models by language models.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. 2023. Introducing our multimodal models.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. 2023. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

Hugo Bronkhorst, Gerrit Roorda, Cor Suhre, and Martin Goedhart. 2020. Logical reasoning in formal and everyday reasoning tasks. *International Journal of Science and Mathematics Education*, 18:1673–1694.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter,

Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems.

AnnaMarie Conner, Laura Singletary, Ryan C. Smith, Patty Anne Wagner, and Richard T. Francisco. 2014. Identifying kinds of reasoning in collective argumentation. *Mathematical Thinking and Learning*, 16:181 – 200.

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

Adnan Darwiche. 2018. Human-level intelligence or animal-like abilities? *Communications of the ACM*, 61(10):56–67.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Igor Douven. 2011. Abduction.

Bradley H Dowden. 2018. Logical reasoning.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Erich Elsen, Augustus Odena, Maxwell Nye, Sağnak Taşırlar, Tri Dao, Curtis Hawthorne, Deepak Moparthi, and Arushi Somani. 2023. Releasing Persimmon-8B.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter v2: Parameter-efficient visual instruction model.

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. 2023. Text-to-audio generation using instruction-tuned llm and latent diffusion model. *arXiv preprint arXiv:2304.13731*.

Usha Goswami. 1991. Analogical reasoning: What develops? a review of research and theory. *Child development*, 62(1):1–22.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR.

Philip N Johnson-Laird. 1999. Deductive reasoning. *Annual review of psychology*, 50(1):109–135.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. 2018. Figureqa: An annotated figure dataset for visual reasoning.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. 2023. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023b. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023c. Seed-bench: Benchmarking multimodal llms with generative comprehension.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023d. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. 2023. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge.

John McCarthy. 2007. From here to human-level ai. *Artificial Intelligence*, 171(18):1174–1182.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. Gpt-4v(ision) system card.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J. Nay, Kshitij Gupta, and Aran Komatsuzaki. 2023. Arb: Advanced reasoning benchmark for large language models.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Generative pretraining in multimodality.

InfiMM Team. 2024. Infimm: Advancing multimodal understanding from flamingo's legacy through diverse llm integration.

InternLM Team. 2023a. Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM`.

The MosaicML NLP Team. 2023b. Introducing mpt-7b: A new standard for open-source, commercially usable llms. `https://www.mosaicml.com/blog/mpt-7b`.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation.

Douglas N Walton. 1990. What is reasoning? what is an argument? *The journal of Philosophy*, 87(8):399–419.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogvlm: Visual expert for pretrained language models.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration.

Fei Yu, Hongbo Zhang, and Benyou Wang. 2023a. Nature language reasoning, a survey. *arXiv preprint arXiv:2303.14725*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023b. Mm-vet: Evaluating large multimodal models for integrated capabilities.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019a. From recognition to cognition: Visual commonsense reasoning.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019b. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2023a. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023b. Llama-adapter: Efficient fine-tuning of language models with zero-init attention.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023c. Multimodal chain-of-thought reasoning in language models.

Bo Zhao, Boya Wu, and Tiejun Huang. 2023a. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023b. Mmicl: Empowering vision-language model with multi-modal in-context learning.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023b. Minigpt-4: Enhancing vision-language understanding with advanced large language models.

Yongshuo Zong, Tingyang Yu, Bingchen Zhao, Ruchika Chavhan, and Timothy Hospedales. 2023. Fool your (vision and) language model with embarrassingly simple permutations.

Appendix

A Detailed reasoning categorizations

Deductive reasoning derives new conclusions from established premises (Johnson-Laird, 1999), ensuring that the steps of inference align with established logical rules. To illustrate, consider the deductive example presented on the right of Figure 1: the premises include observations as "snow is presented in image", "soil is revealed after snow melting, looks like crack", and "crack is expanding". From these premises, the deductive conclusion drawn from premises is "current season is winter, after winter it will be spring". Deductive reasoning capability is vital for MLLMs in various domains. This encompasses automatic fact-checking of multi-modal information and multi-modal legal reasoning for interpreting legal documents, among other applications.

Abductive reasoning determines the most plausible explanation, grounded in common sense for a specific set of observations (Douven, 2011). This form of reasoning is often viewed as the converse of deductive reasoning. in the abductive scenario illustrated in Figure 1, the observation is "a person is cutting an onion while wearing a helmet". Given the commonsense knowledge that "Onions can release compounds causing eyes irritation", the most plausible explanation for the question is "eye protection". The capability of abductive reasoning extends to causal inference in complex systems. It can be applied, but is not limited to, inferring public sentiment from economic data and news, or predicting trends from text, images, and videos.

Analogical reasoning facilitates the transfer of knowledge from known instances to analogous situations (Goswami, 1991). In the example illustrated in Figure 1, the first image demonstrates a proposition that the naming convention is a play on words involving depth. The second and third images should adhere to a similar pattern. Specifically, while the individual in the second image is facing east, the person in the third image faces west, suggesting that his name should logically be "Westface". The capability for analogical reasoning is pivotal in comparative analysis, which constitutes a fundamental aspect of in-context learning.

B Examples with varied complexities

We presents a curated set of examples from our dataset in Figure 5, varying in reasoning complexity, alongside corresponding responses from InfiMM-v1 and GPT-4V.

C MLLM reasoning evaluation benchmarks comparison

We compare our proposed benchmark with prevailing MLLM reasoning evaluation benchmark in Table 5.

Table 5: MLLM reasoning evaluation benchmarks comparison.

Dataset	Domains	Source	# Samples	Answer Format	Reasoning Steps
ScienceQA (Lu et al., 2022)	Natural science, social science, and language science.	Collected from online learning platform.	10332 questions	Multi-Choice	90% with explanations, not used in evaluation.
MathVista (Lu et al., 2023)	Mathematical reasoning with visual contexts.	Aggregated from public datasets and manually collected	6141 questions	55% Multi-Choice, and 44% free-form	Partial contain explanation, not used in evaluation.
MM-Vet (Yu et al., 2023b)	6 VL abilities, including reasoning	Manual collected	218 questions	Free-form	None
MMMU (Yue et al., 2023)	College-level subject knowledge.	Manually collected	11.5K questions	94% Multi-Choice, 6% Free-form	17% with Explanation, not used in evaluation.
InfiMM-Eval (Ours)	Open-ended common sense complex multi-modal reasoning.	Manually collected	279 questions	Free-form text	100% with explanation, used in evaluation.

D Counter-intuitive examples

We provide more counter-intuitive examples of InfiMM-Eval in Figure 6.

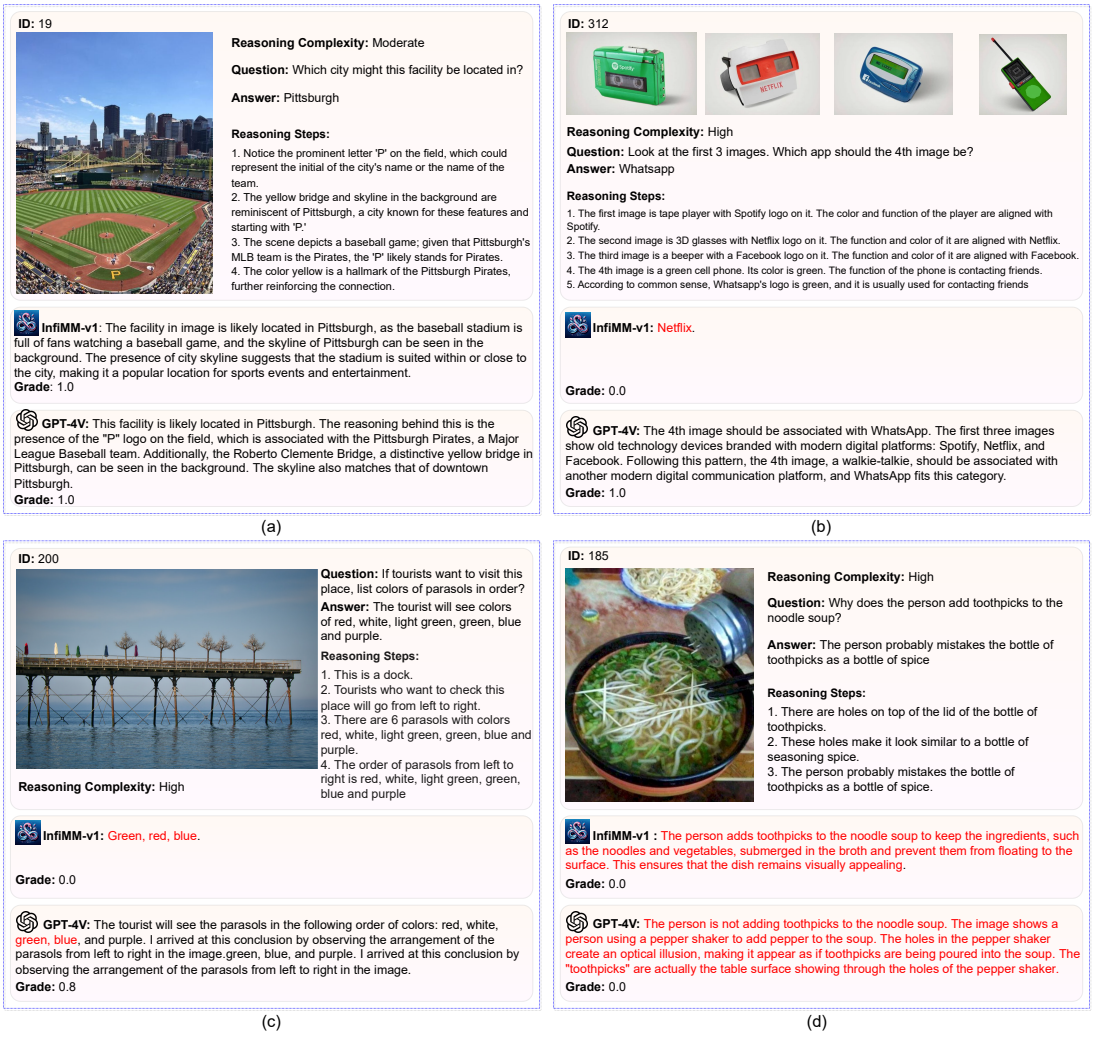

Figure 5: Samples with MLLMs' responses and scores. Hallucinations and errors are highlighted in red.

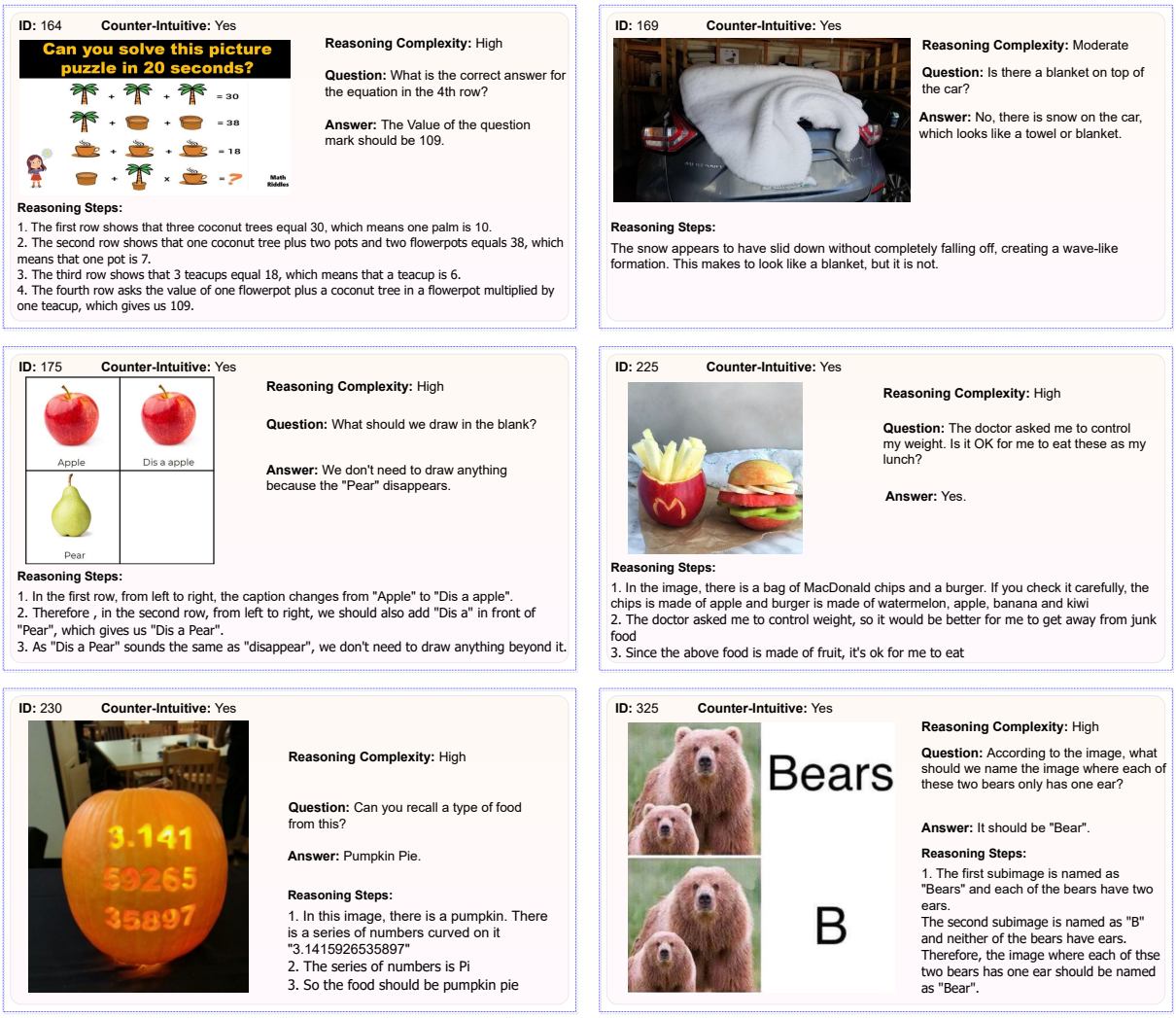

Figure 6: More counter-intuitive examples of InfiMM-Eval.

E Model inference prompts

We list prompts we used for different models in Table 6. For Chain-of-thought prompts, we simply add "Let's think step by step" at the end of the prompt.

F Additional ablation study

In this section, we listed additional ablation studies on InfiMM-Eval.

F.1 Multi-Images as input results

Taking multiple images as input is a crucial capability for MLLMs to do multi-round dialogues and interactive step-by-step reasoning. In this section, we explore current MLLMs' multi-image reasoning capability. We compare MLLM's performance by feeding each image seperately and concatenate multiple images horizontally into a single one. Results are listed below in Table 7.

We select Fuyu-8B, EMU and GPT-4V for comparison since these models should support multiple images as input by design. Fuyu-8B is a pretrained only model, which does not follow instruction very well, thus cannot achieve good results. For EMU, the instruction finetuning data usually do not contain multi-image samples, this could be the reason that there's no evidence of performance improvement. For GPT-4V, there is a substantial drop after concatenating images together. If the trained model internally cuts the image into patches for processing, such as Fuyu-8B, concatenating images into a single image might impact their input patches and lead to worse performance.

Table 6: Prompts used for evaluations of different models. {Image} represents image binary, {Question} stands for the questions.

MLLMs	Inference Parameters	Prompts
GPT-4V	temperature: 0.0 top_p: 0.0 max_tokens: 256	System Prompt: You are a helpful assistant for helping answer questions. Most questions are related to reasoning. User Prompt: Here are a list of image detailed descriptions generated by an AI model: Image 1: {Image} Image 2: {Image} ... Please answer the following question: {Question}
OpenFlamingo-v2	max_new_tokens: 512 num_beams: 3	{image}User: {question} GPT:<answer>
MiniGPT-v2	do_sample: False max_new_tokens: 256	<s>[INST]<Img>{Image} </Img>{Question} [/INST]
Fuyu-8B	max_new_tokens:16	{Image}{Question}
BLIP-2	temperature: 1.0 max_new_tokens: 20	{Image} Question:{Question} Answer:
InternLM-XComposer-VL	temperature: 1.0 max_new_tokens: 1024	<|User|>{Image} {Question}, answer this question <eoh><|Bot|>
InstructBLIP	temperature: 1.0 max_new_tokens: 128	{Image}{Question}
LLaMA-Adapter V2	max_gen_len: 256 temperature: 0.1 top_k: 0.75	Below is an instruction that describes a task. Write a response that appropriately completes the request using a single word or phrase. Instruction: {Image} {Question} Response:
Otter	num_beams:3 max_new_tokens:512	{Image}User: {Question} GPT:
mPLUG-Owl2	max_new_tokens: 256	USER: {Image}{Question} Answer the question using a single word or phrase. ASSISTANT:
IDEFICS-9B-instruct	temperature: 1.0 max_new_tokens:200	User: {image} {Question} Assistant:
Emu	temperature: 1.0 max_new_tokens: 128	System Prompt: You will be presented with an image: [IMG]{Image}[/IMG]. You will be able to see the image after I provide it to you. Please answer my questions based on the given image. <|System Prompt|>USER: {Question} ASSISTANT:
LLaVA-1.5	temperature: 1.0 top_p: 1.0 max_tokens: 256	System Prompt: A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. {Image}...{Image} {Question}
CogVLM-Chat	temperature: 0.8 max_new_tokens: 2048	{Image}{Question}
Qwen-VL-Chat	do_sample: False num_beams: 1 max_new_tokens: 100	<im_start>You are a helpful assistant. <im_end> Picture 1 {Image} Picture 2 {Image} ... {Question}

Table 7: Ablation study results on InfiMM-Eval's subset with multiple images as input. There are 47 samples with multiple images, which contain 27 moderate complexity questions and 20 high complexity questions.

MLLMs	Concatenate	Score (Multi-Img)
Fuyu-8B	Yes	8.21
	No	7.16
EMU	Yes	28.21
	No	27.76
GPT-4V	Yes	57.61
	No	71.19