

# MLLM-Bench: Evaluating Multimodal LLMs with Per-sample Criteria

Wentao Ge\*, Shunian Chen\*, Guiming Hardy Chen\*, Nuo Chen, Junying Chen, Zhihong Chen<sup>†</sup>, Wenya Xie, Shuo Yan, Chenghao Zhu, Ziyue Lin, Dingjie Song, Xidong Wang, Anningzhe Gao, Zhiyi Zhang, Jianquan Li, Xiang Wan, Benyou Wang<sup>†</sup>

Shenzhen Research Institute of Big Data

The Chinese University of Hong Kong, Shenzhen

{wentaoge, shunianchen, guimingchen, zhihongchen}@link.cuhk.edu.cn

wangbenyou@cuhk.edu.cn

## Abstract

Multimodal large language models (MLLMs) have broadened the scope of AI applications. Existing automatic evaluation methodologies for MLLMs are mainly limited in evaluating *objective* queries without considering *real-world* user experiences, inadequately addressing the nuances of creative and associative multimodal tasks. However, the *open-ended* and *subjective* nature of such tasks poses a significant challenge to the evaluation methodology, where it is difficult to define the ground-truth answers for them. To this end, in our paper, we propose a new evaluation paradigm for MLLMs, which is evaluating MLLMs with *per-sample criteria* using potent MLLM as the judge. To validate the feasibility and effectiveness of this paradigm, we design a benchmark, dubbed **MLLM-Bench**, by curating the evaluation samples across six comprehensive cognitive levels. We benchmark 26 popular MLLMs in a pairwise-comparison fashion, showing diverse performance across models. Moreover, the validity of our benchmark manifests itself in reaching 88.02% agreement with human evaluation. We contend that the proposed paradigm explores the potential of MLLMs as effective evaluation tools with the help of per-sample criteria. Our code and data are publicly released at <https://github.com/FreedomIntelligence/MLLM-Bench>.

## 1 Introduction

The field of artificial intelligence has witnessed a paradigm shift with the incorporation of multimodal inputs, particularly the amalgamation of visual and linguistic data, mirroring the complex processing capabilities of the human brain. The development of multimodal large language models (MLLMs) (e.g., MiniGPT-4 (Zhu et al., 2023), LLaVA (Liu et al., 2023b,c), Qwen-VL (Bai et al.,

2023a), Cluade-3-Opus (Anthropic, 2024), and GPT-4V (OpenAI, 2023)) represents a leap towards more sophisticated, context-aware AI systems. These models are increasingly crucial for tasks that demand an understanding of both visual cues and textual content. However, the expansion of capabilities brings forth the challenge of evaluation – how does one accurately measure the effectiveness of a system designed to mimic the inherently subjective and associative processes of human perception?

The predominant evaluation frameworks for MLLMs focus on close-ended queries with clear-cut, correct answers. Such tasks are valuable in quantitating the model performance but do not consider user experience and encapsulate the full spectrum of human-like cognitive tasks that modern MLLMs aim to perform. The realms of creativity, association, and ethical judgment within multimodal contexts, for instance, resist reduction to simple right or wrong answers. However, human evaluation is highly inefficient, while automatically evaluating the MLLMs’ performance on open-ended queries is challenging. This limitation in evaluation methodologies leads to an incomplete understanding of a model’s capabilities and fails to provide insight into how such models might interact with users in real-world scenarios, where answers are often nuanced and context-dependent.

To bridge this gap, we propose to use potent MLLM<sup>1</sup> as the judge with **per-sample criteria** to evaluate MLLMs. To validate this paradigm, we develop a comprehensive benchmarking suite, named **MLLM-Bench**, including 42 distinct aspects of MLLM functionality, distributed among six critical levels of capability: *perception*, *understanding*, *applying*, *analyzing*, *evaluating*, and *creation*, inspired by Bloom’s Taxonomy (Krathwohl, 2002).

\*The first three authors contribute to this work equally.

<sup>†</sup>Zhihong and Benyou are the corresponding authors.

<sup>1</sup>We adopt GPT-4-Turbo as the judge for main experiments, while we also show the result using Cluade-3-Opus.



**Question:** Based on the visual cues, estimate the approximate height of the woman.

**Answer 1 (LLaVA-v1.5-13B):**

The woman in the image is relatively short, standing at around 3 feet tall.

**Answer 2 (LVIS-instruct4v-LLaVA-7B):**

The woman is standing between two tall men, and based on the visual cues, she appears to be quite short in comparison. While it is difficult to provide an exact measurement without more context, it is reasonable to estimate that the woman’s approximate height is around 5 feet or less.

**Criteria**

1. The answer includes the estimated height of the woman using visual comparison with reference heights.
2. Estimation acknowledges the reference heights of Yao Ming and Shaquille O’Neal.
3. The height range of the woman should be 165cm to 175cm.

**GPT-4-Turbo Judgment**

Upon reviewing the provided responses in relation to the visual cues from the image and the given evaluation criteria:

**Answer 1** suggests that the woman is “around 3 feet tall.” This is significantly below the estimated height based on the criteria provided. Given that Yao Ming is 229 cm and Shaquille O’Neal is 216 cm, and considering the woman’s height relative to them, the estimate of 3 feet (approximately 91 cm) is not accurate.

**Answer 2** estimates the woman’s height to be “around 5 feet or less”. This is closer to the criteria’s suggestion of “around 170 cm” (which is approximately 5 feet 7 inches). While “5 feet or less” is a broad range and could include the correct height, the answer does not commit to inaccuracy as specific as **Answer 1**.

Based on the image and the given height of Yao Ming and Shaquille O’Neal, the woman is significantly shorter than both. However, she is not as short as 3 feet, as suggested by **Answer 1**. **Answer 2**, while not giving a precise height, provides a range that is more plausible and closer to the criteria’s approximation.

Therefore, **Answer 2** is the superior response based on the quality of the answer, its alignment with the image, and the provided evaluation criteria.

Table 1: Pair-wise evaluation using per-sample criteria for MLLMs. The per-sample criteria is only available to the GPT-4-Turbo judge and are not accessible to evaluated MLLMs such as LLaVA-v1.5-13B and LVIS-instruct4v-LLaVA-7B for answer generation. **GPT-4-Turbo Judgment** shows the judgment rationale of GPT-4-Turbo and is for demonstration purposes. Our experiment only requires the judge to output the judgment result (see details in Section 5.1).

Notably, rather than providing a standard answer to a posed question, we offer one to three specific *evaluation criteria for each question*. These criteria are then applied in conjunction with the posed question and the model’s response to assess performance more accurately (see an example in Table 1).

The contributions of this paper are two-fold: 1) **A new paradigm to evaluate MLLMs** utilizing **powerful MLLMs** along with **per-sample criteria**, which shifts from closed-form-answer evaluations for MLLMs to a flexible, criteria-based approach, particularly suited for open-ended tasks. It recognizes and acknowledges a spectrum of valid responses and evaluates the answer quality based on how well they align with these criteria, assessing models beyond the single “correct” answer limitation. 2) **A benchmark dataset** with a comprehensive taxonomy that categorizes tasks and scenarios for MLLMs, with a strong emphasis on ethical considerations. In the benchmark, we conduct a **systematic benchmarking of existing MLLMs**. The benchmarking is unique since it might be well-aligned to user experience in real-world applications where questions are usually without unique groundtruth.

## 2 Background: Evaluations for MLLMs

It is challenging to comprehensively assess the capabilities of MLLMs. Current benchmarks primar-

ily fall into several categories: (1) Multiple-choice questions (evaluating the perception and cognition abilities of MLLMs): MME (Fu et al., 2023), SEED (Li et al., 2023a), and TouchStone (Bai et al., 2023b); (2) Arena-like evaluation (user-based evaluation of different capabilities): LVLm-eHub (Xu et al., 2023), VisIT-Bench (Bitton et al., 2023); (3) Hallucination assessment (focusing on a key issue currently faced by MLLMs - hallucinations): POPE (Li et al., 2023e) and HallusionBench (Liu et al., 2023a). The works most related to us are (i) MMBench (Liu et al., 2023d) and MM-Vet (Yu et al., 2023), using GPT-4 as the evaluator to quantitatively measure the performance of different MLLMs; (ii) a concurrent work (Zhang et al., 2023) uses GPT-4V to evaluate text-to-image generation.

**Open- and Closed-ended Benchmarks** Existing benchmarks are categorized into either open-ended or closed-ended with different evaluation approaches. Although close-ended benchmarks such as MMLU (Hendrycks et al., 2020), C-Eval (Huang et al., 2023) for LLMs and MME (Fu et al., 2023) and SEED (Li et al., 2023a) for MLLMs are convenient to evaluate on, they often suffer from data contamination issue. The results of close-ended benchmarks are especially for proprietary LLMs whose training data are all *in-house*. On the other hand, open-ended benchmarks (e.g., MT-Bench (Zheng et al., 2023) and Alpaca-Eval (Li et al., 2023d))

test models via free-form generation, which is more consistent with real-world use cases and relies heavily on LLMs’ generation ability. The issue of data contamination in open-ended benchmarks is less severe since there are no standard answers, and such contamination offers minimal assistance in benchmark hacking.

**LLMs for MLLM Evaluation** MMBench (Liu et al., 2023d), TouchStone (Bai et al., 2023b), and MM-Vet (Yu et al., 2023) employ LLM-based evaluation frameworks, leveraging advanced LLMs (e.g., GPT-4) to assess MLLMs. This approach encounters significant limitations due to the inability of text-only LLMs to perceive visual contexts directly. The idea of adopting GPT-4-Turbo, a potent MLLM, directly as a judge in this paper, marks a significant advancement in the field.

### 3 Motivation of MLLM Evaluation with Per-sample Criteria

#### 3.1 Motivations

**Why MLLM-as-the-judge evaluation needs additional criteria?** While potent MLLMs are potential evaluators, their assessment outcomes may not always align perfectly with factual accuracy or human standards. This discrepancy highlights the necessity for a more nuanced approach to evaluation: per-sample criteria. Per-sample criteria are designed to provide specific benchmarks and guidelines for each assessment task, aiding MLLM judge for MLLM evaluation. This approach is particularly valuable for tasks where the judge’s capabilities might fall short, ensuring that evaluations remain robust even in areas of potential weakness. Unlike a single reference answer, per-sample criteria afford a broader and more flexible basis for assessment, making them ideally suited for evaluating open-ended questions. As we adopt GPT-4-Turbo as the judge in this paper, an example where GPT-4-Turbo falls short is shown in Appendix E.1.

**Why criteria should be sample-specific?** Different samples present unique challenges and requirements, making a one-size-fits-all approach to evaluation inadequate. Table 2 exemplifies the critical per-sample criteria essential for the nuanced evaluation of multi-modal large language models (MLLMs), especially in *Soft Reference*, *Range*, and *Evaluation Guideline* categories. These criteria underscore the need for evaluations that adapt to the task’s context and intricacies, showcasing the

limitations of previous methodologies in comprehensively assessing MLLM capabilities.

Criteria Type	Description	Example criteria (images omitted)
Exact Reference	Necessitates precise answers, such as accurately determining an item’s location.	<i>Answer specifies cactus location as row 2, column 6.</i>
Soft Reference	Requires identifying and translating contextually relevant texts, demanding linguistic adeptness and current knowledge.	<i>Identify and translate text related to the novel coronavirus pneumonia in Wuhan.</i>
Range	Allows for variability within defined limits, like estimating a person’s height, introducing flexibility.	<i>The height range of the woman should be 165cm to 175cm.</i>
Evaluation Guideline	Involves interpretative analysis to evaluate complex situational effects on responses.	<i>Assess if the answer considers the environment’s condition, like submerged paths.</i>

Table 2: Typical per-sample criteria. The criteria are sampled from this benchmark.

#### 3.2 Benefits of Per-sample Criteria

##### Generalization of Referenced-based Evaluation

For questions with *objective* answers, the criteria offer specific reference answers, as demonstrated in the first, third, and fifth samples in Table 4. In cases where the instructions require a *subjective* description or yield ambiguous answers, the criteria supply essential information for formulating responses, exemplified by the second, fourth, and sixth samples in the same table. This approach enables the evaluation model (GPT-4-Turbo) to assess the quality of outputs using a well-defined standard.

**Mitigation of Data Contamination** One of the advantages of our per-sample criteria is its potential to alleviate data contamination problems. While we will continuously update our dataset, there is no guarantee that the samples have not been seen or used. To this end, we choose to withhold the per-sample criteria when publishing our dataset. Namely, the released version will only include the images and instructions. We believe that this approach significantly reduces the risk of contamination since even if models have been exposed to the images, they do not necessarily generate desired responses.

## 4 MLLM-Bench Dataset

### 4.1 Taxonomy of Capabilities

Due to the absence of a standardized framework for categorizing the capabilities of multimodal large language models, and acknowledging that vision-language models emulate human cognitive processes to a certain extent, we have chosen to adopt the revised Bloom’s Taxonomy (Krathwohl, 2002) as the framework for this benchmark. In reference to Bloom’s Taxonomy, we manually conclude 42 capabilities of MLLMs across a hierarchy spanning six cognitive levels and create 10 questions for each capability. The six capability levels are shown below in Table 3. For each of the capabilities, we create 10 questions, resulting in a total of 420 image-instruction pairs, see details in Appendix A.

### 4.2 Data Annotation

#### 4.2.1 Data Annotators

We have recruited six volunteers, all of whom are undergraduates, graduate students, or research assistants at a university with an all-English curriculum. Each volunteer is paid according to the local salary (i.e., equivalent to roughly 10 dollars per hour) and tasked with gathering data pertinent to a distinct capability level, thereby guaranteeing consistency within each specific category. The data collection phase spanned a duration of two weeks; they can complete it whenever convenient. Before data collection and annotation, they are instructed to follow a guideline. See details in Appendix B.1.

#### 4.2.2 Pipeline

The data collection and annotation module comprises four stages:

**I: Image Collection:** Volunteers gather the most recent images that are pertinent to the capabilities under examination, either from social networks or by capturing them in real life.

**II: Instruction Construction:** Utilizing GPT-4-Turbo, volunteers craft assessment instructions that are congruent with the requisite model capabilities, the context of the collected images, and the manually written prompts. This stage also includes a consistency check between the instructions and the corresponding images.

**III: Question Type Annotation:** Recognizing that our benchmark evaluates model performance on both traditional closed-ended questions and more exploratory open-ended tasks, volunteers are

Level	#Samples	Description
Perception	70	MLLMs retrieve information from multimodal inputs, using skills like object recognition and OCR.
Understanding	110	MLLMs process perceived information to construct meaning, comprehending and interpreting data contextually.
Applying	60	MLLMs apply knowledge to similar situations, such as using text-based knowledge to interpret images as in medical imaging.
Analyzing	120	MLLMs break down information to explore relationships, performing tasks like attribute comparison or causal reasoning.
Evaluation	40	MLLMs make judgments based on criteria and standards, like assessing image quality or discerning content authenticity.
Creation	20	MLLMs synthesize information to generate new content, from visual storytelling to coding with vision.

Table 3: Taxonomy of MLLM-Bench including examples at six cognitive levels.

required to annotate the type of each question. This includes categorizing them as *open-ended*, *closed-ended*, or *compound*. Compound questions contain elements of both open-ended and closed-ended queries, offering a more comprehensive challenge to the models being tested.

**IV: Per-Sample Criteria Annotation:** Volunteers are asked to provide evaluation criteria based on the image, instructions, capability requirements, and question type for each item of data. These criteria include *exact references*, *soft references*, *acceptable ranges*, and *evaluation guidelines*.

#### 4.2.3 Guidelines for Annotators

**Data Protocol** As illustrated in Table 3, each entry in the MLLM-Bench dataset comprises three key components: 1) a contemporary **image** with a friendly license, 2) a **question** posed as it would naturally arise in real-world situations and 3) 1-3 customized **per-sample criteria** that offer guidelines specifically designed to complement the capabilities of GPT-4-Turbo, thus enabling a more logical and scientifically sound evaluation.

**Guideline for the Data Annotation** The guideline for data annotation emphasizes the importance of using recent images to avoid data leak-



age<sup>2</sup>, sourcing data from publicly licensed platforms like Twitter or direct captures with clear copyright status, ensuring image clarity while accommodating real-world quality variance, maintaining impartiality by excluding sensitive content, and promoting diversity in response formats to reflect complex real-world interactions. See details in Appendix B.2.

### 4.3 Data Quality Control

We employ a two-step protocol to validate the data: *cross-review* and *expert verification*.

**Cross-review** Upon completion of data collection, the six volunteers are required to review each other’s work following the criteria in Table 10 in Appendix B.2. Samples that did not meet the criteria were discarded and replaced to meet the required quota for each category. We conducted three rounds of cross-review, as shown in Figure 1.

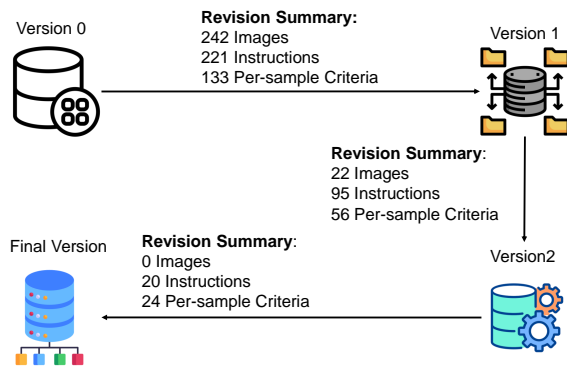


Figure 1: Number of revised samples during the cross-review process.

**Expert Verification** We recruited an experienced volunteer with expertise in data evaluation to inspect each sample for quality assurance. Low-quality samples are recollected using a process similar to cross-verification.

## 5 Benchmarking

### 5.1 Evaluation Protocols

**Evaluation Settings** Our evaluations are conducted in a *pairwise voting* manner. As illustrated in Table 1, we adopt the superb GPT-4-Turbo<sup>3</sup> as the judge to vote for which answer is better given an image, a question, and a pair of answers. Each

<sup>2</sup>To mitigate the possible data leakage issue, one way is to continuously maintain and expand the current benchmark dataset in real-time.

<sup>3</sup>We use gpt-4-turbo-2024-04-09 for evaluation.

pair of answers consists of a benchmarked answer from an MLLM to be judged and an answer from the anchor model LaVA-v1.5-13B, which is a moderately strong MLLM. We compute the number of win/tie/lose outcomes of a benchmarked model over the anchor model. The primary reason for selecting LLaVA-v1.5-13B as the anchor model is its popularity as a baseline at the time of our experiments. This choice provides intuitive and direct insights into model effectiveness, given the community’s familiarity with LLaVA-v1.5-13B. Additionally, choosing a model that is neither too strong nor too weak helps in effectively distinguishing performance differences among the tested models.

One might be concerned that a win rate over an anchor model relies heavily on the choice of anchor model. As a result, we also adopt a **bootstrap version of the Elo rating system** as used in Chatbot Arena (Zheng et al., 2023) to rank all benchmark models.

Previous works (Wang et al., 2023b; Wu and Aji, 2023; Chen et al., 2024a) investigate the positional bias in text-only evaluation. In this work, we also evidence that such bias also exists in multi-modal evaluation using GPT-4-Turbo (see Section 6.4), albeit subtle. To mitigate positional bias, we shuffle the position of each paired sample for all settings.

**Evaluation Prompts** The evaluation criteria are based on the responses’ quality, relevance to the image, as well as the given per-sample criteria. The evaluator is to choose which of the two answers is superior, except in two specific situations where a decision cannot be made: **1)** if the question is close-ended and both answers are equally correct or incorrect, or **2)** if both answers have significant factual inaccuracies or ethical concerns. In such cases, the evaluator can indicate their inability to decide, specifying the situation. In the last line, the judge should output a concise declaration of which answer is better or an indication that a decision cannot be made, citing the appropriate reason. Figure 6 in Appendix E shows the evaluation prompt.

**Benchmarked MLLMs** We select 26 MLLMs to be evaluated based on two factors: popularity and architectural diversity. Primarily, the models chosen are some of the most widely used MLLMs at the time of this study, substantiated by their number of GitHub stars. Secondly, these models offer a broad spectrum of MLLM architectures, allowing us to conduct a comprehensive and comparative study. Among these models, **GPT-family, Claude-**







Capability	Image	Sample Questions	Criteria
Food Recognition		What type of mushroom is it in the image? Is it edible?	<b>1.</b> Identify mushroom as <i>galerina marginata</i> . <b>2.</b> State the mushroom is not edible. <b>3.</b> Ensure the answer is based on visual characteristics of the mushroom in the image.
Attribute Recognition		Please describe the feature of the animal as detail as possible to help other people to recognize it.	<b>1.</b> Mention toucans' large, colorful bills as a distinctive feature. <b>2.</b> Describe additional attributes: body color, size, and any unique markings. <b>3.</b> Include the setting or behavior if it aids in recognition.
Object Localization		The image is of dimension 700*1050, assume the bottom left corner to be the origin of coordinates (0,0), the coordinates of the top left corner to be (0,1050) and the coordinates of the bottom right corner to be (700,0). Please give the approximate coordinates of the spoon tail.	<b>1.</b> Answer should include coordinates close to the reference: (200,315), the closer the better.
Function Reasoning		The image depicts a bike that has been a part of my collection for quite some time. If I were to consider riding it, could you please advise me on the components required to restore it to working condition?	<b>1.</b> Identify missing/damaged components: front tire, chain, pedal, handle bar, seat. <b>2.</b> Explain the function of each identified component for bike restoration.
Fake Image Detection		Determine if this photo has been edited and provide a detailed explanation.	<b>1.</b> Confirm the photo is edited by identifying the added squirrel. <b>2.</b> Evaluate the explanation of how the squirrel's integration indicates editing. <b>3.</b> Assess the consistency of lighting and shadows related to the squirrel.
Coding Capability with Vision		The image depicts a game. You need to analyze the gameplay and use code to implement an identical game.	<b>1.</b> Identify game as sliding puzzle with numbered tiles; mention grid and tile combination mechanics. <b>2.</b> Describe the gameplay objective to combine tiles for a higher-numbered tile. <b>3.</b> Include code implementation reflecting game rules and sliding tile mechanism.

Table 4: Data samples in MLLM-Bench, which are presented from top to bottom across six capability levels: Perception, Understanding, Applying, Analyzing, Evaluation, and Creation.

**family**, and **Gemini-1.5-family** represent cutting-edge proprietary solutions, leveraging state-of-the-art techniques for multimodal understanding and generation. For open-sourced MLLMs, the LLaVA-series and their variants (LVIS, MiniGPT-v2, Yi-VL, and ALLaVA provide robust solutions for multimodal adaptation and generalization. Qwen-VL, Qwen-2-VL, Cheetor, and mPLUG-Owl2 use query tokens as the bridge between visual and textual representations. Similarly, Pixtral shows strong performance on multimodal tasks, while maintaining state-of-the-art performance on text-only benchmarks. It supports multiple images in the long context window of 128k tokens. LWM, with its ability to support a 1 million token context, pushes the boundaries of context handling in MLLMs. SEED-LLaMA uses a VQ-based encoder, while Fuyu adopts a single linear layer to link two modalities. Lastly, Phi-3-Medium strikes a balance between efficiency and accuracy, offering strong performance in multimodal tasks. Models are detailed in Appendix D.

## 5.2 Evaluation Results

Table 5 presents the results of benchmarked models. The outcomes are quantified as the number of wins, ties, and losses for each model relative to LLaVA-v1.5-13B, with the models sorted by their total win rates in descending order. The last two columns list the Elo ratings and the rankings of models with respect to their Elo scores.

Overall, closed-source models demonstrate a leading advantage on the leaderboard, with some closed-source models such as Claude-3-5-sonnet and the Gemini-1.5 series showing outstanding performance. However, at the same time, the gap between open-source and closed-source models is narrowing. Leading open-source models like Pixtral and the Qwen-2-VL series are exhibiting levels close to those of closed-source models. Nevertheless, we must acknowledge that there are significant variations in the performance of open-source models, and further efforts are needed in their pursuit to catch up with closed-source models.

In the meanwhile, we calculate the Spearman coefficient between the anchor-base method and

Models	Perception	Understanding	Applying	Analyzing	Evaluation	Creation	Win Rates	Elo Ratings	Elo Rankings
Claude-3-5-sonnet	67/0/3	106/4/0	56/2/2	90/8/2	38/1/1	39/0/1	0.94	1438	1
Gemini-1.5-pro	69/1/0	101/8/1	53/5/2	88/6/6	38/1/1	39/0/1	0.92	1368	2
Gemini-1.5-flash	67/2/1	99/7/4	49/8/3	87/9/4	40/0/0	36/4/0	0.90	1346	3
Claude-3-opus	56/13/1	98/9/3	45/11/4	83/14/3	33/5/2	33/6/1	0.83	1267	4
GPT-4o-mini	61/6/3	90/9/11	45/10/5	77/10/13	32/2/6	38/2/0	0.82	1189	5
GPT-4-vision-preview	56/10/4	92/9/9	40/17/3	84/11/5	32/2/6	33/5/2	0.80	1106	9
Pixtral-12B	55/8/7	88/8/14	46/7/7	72/14/14	32/2/6	36/1/3	0.78	1170	6
Qwen-2-VL-72B	49/5/16	78/15/17	40/10/10	72/7/21	32/2/6	32/1/7	0.72	1114	8
Qwen-2-VL-7B	56/10/4	67/24/19	39/13/8	64/12/24	23/6/11	32/6/2	0.67	1065	12
LLaVA-v1.6-34B	46/17/7	78/22/10	36/15/9	61/28/11	33/3/4	24/10/6	0.66	1104	10
LLaMA-3.2-Vision-11B	47/13/10	74/18/18	33/13/14	66/15/19	26/8/6	25/5/10	0.65	1086	11
LLaVA-v1.6-Vicuna-13B	40/21/9	65/33/12	35/19/6	51/26/23	33/5/2	27/9/4	0.60	1127	7
LLaVA-v1.5-13B ( <i>anchor</i> )							0.50	/	/
LLaVA-v1.6-Vicuna-7B	31/25/14	56/37/17	26/23/11	40/31/29	22/10/8	19/10/11	0.46	1018	14
ALLaVA-3B-Longer	22/21/27	57/30/23	23/17/20	44/30/26	16/10/14	17/12/11	0.43	1023	13
Gemini-1.0-Pro	45/10/15	36/35/39	24/19/17	33/28/39	9/8/23	16/8/16	0.39	961	15
Qwen-VL-Chat	34/22/14	38/36/36	26/18/16	35/29/36	15/6/19	9/12/19	0.37	913	17
LVIS	22/28/20	32/39/39	11/27/22	33/36/31	14/9/17	9/16/15	0.29	921	16
mPLUG-Owl2	16/24/30	30/34/46	17/17/26	23/38/39	15/8/17	11/14/15	0.27	889	19
LLaVA-v1.5-7B	19/22/29	27/47/36	13/29/18	21/43/36	9/14/17	8/13/19	0.23	910	18
Phi-3-Medium	8/12/50	11/19/80	14/14/32	26/18/56	6/9/25	17/3/20	0.20	755	22
MiniGPT-v2	12/25/33	24/32/54	11/25/24	17/38/45	9/9/22	6/6/28	0.19	837	20
Cheetor	12/20/38	7/27/76	10/22/28	16/23/61	4/4/32	3/4/33	0.12	761	21
SEED-LLaMA	16/15/39	5/25/80	10/21/29	7/25/68	3/7/30	3/3/34	0.10	641	24
Yi-VL-6B	4/17/49	8/22/80	5/27/28	5/29/66	3/9/28	3/9/28	0.07	739	23
Fuyu-8B	7/19/44	7/27/76	6/14/40	4/22/74	3/7/30	0/6/34	0.06	638	25
LWM	2/18/50	5/15/90	4/21/35	2/18/80	3/2/35	2/6/32	0.04	618	26

Table 5: Number of wins/ties/loses for each model over **LLaVA-v1.5-13B** (*anchor*) on each level. Models are sorted by overall win rates in descending order.

the Elo method, achieving a value of 98.02%. This result demonstrates the robustness of our approach.

## 6 Validating Evaluation Paradigm and MLLM-Bench

We validate our evaluation paradigm through five critical dimensions: alignment with human judgments(Section 6.1), the impact of per-sample criteria (Section 6.2), robustness against test variations(Section 6.3), bias analysis(Section 6.4), and the consistency of different MLLM judges (Section 6.5).

### 6.1 Alignment with Human Evaluation

**Settings** For this alignment experiment, we recruited four undergraduate volunteers<sup>4</sup>, all of whom are students at a university with an all-English curriculum. For volunteers participating in human evaluations, we sample 252 evaluation items. Each item in the evaluation set comprises a question associated with an image and two answers generated by models: one from the model under evaluation and the other from the anchor model LLaVA-v1.5-13B. We present the details of sample and model selection in Appendix F.1.

To check the consistency between human evaluators and GPT-4-Turbo, we calculate the agreement

<sup>4</sup>Each volunteer is paid according to the local salary (i.e., equivalent to roughly 10 dollars per hour).

between each individual’s results as well as the aggregated results of human evaluators and GPT-4-Turbo. The aggregated results are derived from a majority vote among four human evaluators. This step combines individual judgments into a consensus and addresses situations like indecision or equal preference. If selections for the evaluated model and the anchor model are equal, or if more evaluators choose "*unable to decide*" than any specific model, the outcome is classified as "*unable to decide*".

**Results** The last row of Table 6 shows a high alignment between human evaluators and GPT-4-Turbo, with the agreement between individual evaluators and GPT-4-Turbo reaching 86.59% and the agreement between the aggregated results of human evaluators and GPT-4-Turbo reaching 88.02%. This high degree of alignment reaches the same level of agreement among human evaluators. The results validate the premise that GPT-4-Turbo can effectively mirror human judgment in the setting of MLLM evaluation, confirming GPT-4-Turbo as a reliable and effective tool in the evaluation process.

### 6.2 Ablation on Per-sample Criteria

**Settings** To further demonstrate the superiority of our method, we include some baselines for comparison. **GPT-4 + detection**<sup>5</sup> (w/ or w/o criteria): For

<sup>5</sup>Details of bounding box generation are in Appendix F.2.

Methods	Judge 1	Judge 2	Judge 3	Judge 4	Aggregated
GPT-4 + <i>det</i> w/o Criteria	79.10%	78.20%	71.54%	72.48%	75.37%
GPT-4 + <i>det</i> w/ Criteria	82.71%	78.79%	72.87%	80.18%	81.20%
GPT-4 + <i>cap</i> w/o Criteria	80.00%	81.72%	74.59%	82.14%	81.15%
GPT-4 + <i>cap</i> w/ Criteria	78.72%	80.32%	74.58%	80.98%	80.53%
GPT-4 + <i>det</i> + <i>cap</i> w/o Criteria	81.12%	83.42%	75.27%	84.12%	83.00%
GPT-4 + <i>det</i> + <i>cap</i> w/ Criteria	80.41%	82.29%	74.86%	83.54%	83.16%
GPT-4-Turbo w/o Criteria	81.62%	82.22%	75.42%	82.05%	82.80%
<b>GPT-4-Turbo w/ Criteria (ours)</b>	<b>85.37%</b>	<b>86.59%</b>	<b>76.88%</b>	<b>82.55%</b>	<b>88.02%</b>

Table 6: Alignment between human evaluation and different evaluation methods. *det*: detection; *cap*: caption. The **Aggregated** column aggregates the results of 4 human evaluators.

each image, we adopt Detic (Zhou et al., 2022) generate top-10 BBoxes ranked by confidence score. Then we feed GPT-4 with BBox coordinates for evaluation. **GPT-4 + caption**<sup>6</sup> (w/ or w/o criteria): For each image, we generate a detailed caption using GPT-4-Turbo. Then we feed GPT-4 with the captions for evaluation. **GPT-4 + detection + caption (w/ or w/o criteria)**: For each image, top-10 BBox coordinates and a detailed caption are sent to GPT-4 in a single prompt for evaluation. **GPT-4-Turbo and GPT-4-Turbo + criteria (ours)**: Identical setting as in Section 6.1. Prompts for GPT-4 evaluation are in Appendix F.4.

**Results** Table 6 summarizes the results. Per-sample criteria bring more than 5% overall agreement increase on GPT-4+*det* and GPT-4-Turbo evaluation, demonstrating the effectiveness of our proposed method. Evaluation using GPT-4 with compound information (*det* + *cap*) outperforms using either of them, suggesting that these two methods introduce distinct information to GPT-4. We also relate our work to a recent work (Fu et al., 2024), which finds that GPT-4 performs well on some MLLM benchmarks with dense captions as input. Nevertheless, GPT-4-Turbo-as-a-judge still outperforms GPT-4-as-a-judge under our setting in terms of agreement with humans, which demonstrates the necessity of adopting a vision model for evaluation in MLLM-Bench.

### 6.3 Robustness Analysis

To demonstrate the robustness of our evaluation paradigm, we conduct complementary analyses from two perspectives: evaluation consistency and prompt perturbation. All experiments maintain identical settings as described in Section 5.1.

<sup>6</sup>The prompt for caption generation with GPT-4-Turbo is in Appendix F.3.

**Evaluation Consistency** We select three models and test them three times using GPT-4-Turbo. Table 7 presents the average win rates with standard deviations across cognitive levels. The results in repeated measurements confirm high inter-evaluation consistency. This reproducibility strengthens the reliability of adopting GPT-4-Turbo as the judge in our evaluation paradigm.

**Prompt Perturbation** We examine prompt sensitivity through two controlled experiments: (1) Semantic-preserving prompt restructuring via GPT-4-Turbo, and (2) Chain-of-Thought augmentation requiring post-selection reasoning. The results show that the maximum total score variation across settings is merely 2%. This demonstrates the criteria’s inherent stability against prompt variations. See Appendix F.5 for detailed prompts and experiment results.

### 6.4 Bias Analysis

To address potential concerns about biases in our evaluation paradigm, we conduct analyses on two critical biases: positional bias and length bias.

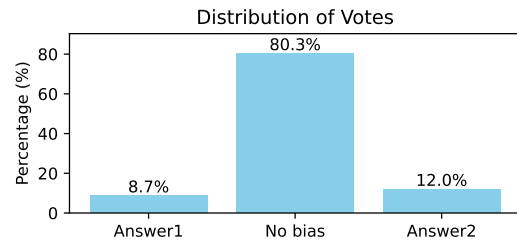


Figure 2: Distribution of votes from GPT-4-Turbo.

**Positional Bias** We examine whether answer ordering affects judgments by comparing votes between original and shuffled answer pairs. Specifically, if GPT-4-Turbo has consistent votes when orders of a pair of answers are shuffled, then we



Models	Perception	Understanding	Applying	Analyzing	Evaluation	Creation	Total
LLaVA-v1.6-34B	0.62 (0.05)	0.68 (0.03)	0.59 (0.01)	0.61 (0.00)	0.82 (0.02)	0.65 (0.04)	0.65 (0.01)
Openflamingo	0.12 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.03 (0.02)	0.02 (0.01)	0.04 (0.01)
Yi-VL-6B	0.05 (0.01)	0.06 (0.01)	0.11 (0.02)	0.06 (0.01)	0.07 (0.01)	0.08 (0.01)	0.07 (0.00)

Table 7: Consistency analysis through triplicate evaluations. Values show mean win rates with standard deviations.

Models	Perception	Understanding	Applying	Analyzing	Evaluation	Creation	Win Rates
Qwen-VL-Chat	31/21/18	37/24/49	25/20/15	35/25/40	9/9/22	11/11/18	0.35
LVIS	26/17/27	30/35/45	15/26/19	34/28/38	11/14/15	11/12/17	0.30
mPLUG-Owl2	18/18/34	28/24/58	13/24/23	19/27/54	11/8/21	8/9/23	0.23
MiniGPT-v2	20/16/34	23/28/59	11/23/26	23/22/55	11/6/23	4/9/27	0.22
SEED-LLaMA	17/12/41	10/21/79	13/17/30	11/20/69	7/4/29	1/7/32	0.14
InstructBLIP	12/18/40	13/14/83	10/11/39	11/22/67	6/7/27	2/7/31	0.13
Fuyu-8B	9/16/45	4/20/86	5/11/44	6/14/80	2/1/37	0/1/39	0.06
kosmos2	5/19/46	3/11/96	2/13/45	2/23/75	2/1/37	0/5/35	0.03
BLIP2	4/11/55	2/11/97	4/9/47	1/11/88	1/2/37	0/1/39	0.03

Table 8: Number of wins/ties/loses for each model on each level, adopting **LLaVA-v1.5-13B** as the anchor and **Claude-3-Opus** as the judge. Models are sorted by overall win rates in descending order.

mark the vote as unbiased (No bias). Otherwise, we mark it as positionally biased towards the first (Answer 1) or the second answer (Answer 2). As shown in Figure 2, 80.3% of judgments remain consistent regardless of answer order, demonstrating strong positional robustness.

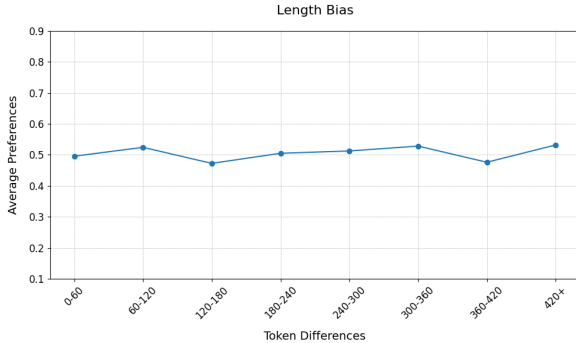


Figure 3: Preference towards different output lengths of GPT-4-Turbo.

**Length Bias** We conducted an experiment to examine GPT-4-Turbo’s preference for output length, as illustrated in Figure 3. The y-axis shows GPT-4-Turbo’s average preference, with 0 indicating a preference for short answers and 1 for long answers. The x-axis displays the token differences between the evaluated answers. The results suggest that GPT-4-Turbo’s preference for output length is negligible.

## 6.5 Effects of Different MLLM Judges

While we use GPT-4-Turbo for our experiments, we aim to show that the framework is designed to

be adaptable and can seamlessly incorporate other potent MLLM models. To prove this, we conduct the same setting in Section 5.1, except that we replace GPT-4-Turbo with Claude-3-opus as the evaluator. We only evaluate a subset of models due to a limited budget. As shown in Table 8, the resulting ranking has a Spearman correlation of 0.95 with GPT-4-Turbo’s voting results, which demonstrates that even if we change the model evaluator, the evaluation results highly align with the original ones. This flexibility allows for the substitution of GPT-4-Turbo with other powerful models, iterating along with the rapidly evolving field of MLLMs.

## 7 Conclusion

In this paper, we propose a new paradigm for MLLM evaluation and present MLLM-Bench, a benchmark for automatically evaluating the MLLMs’ ability on open-ended queries. It is derived from a comprehensive taxonomy, paving the way for a more responsible and conscientious approach to AI development. In MLLM-Bench, instead of providing the judge model with a standard answer, we equip each evaluation sample with evaluation criteria to evaluate open-ended answers reasonably. Experimental analysis on MLLM-Bench shows that evaluation using a potent MLLM using our per-sample criteria strategy aligns better with human than other baselines. We hope that MLLM-Bench can introduce more insights to MLLM evaluation and development.

## Acknowledgements

This work was supported by the Shenzhen Science and Technology Program (JCYJ20220818103001002), Shenzhen Doctoral Startup Funding (RCBS20221008093330065), Tianyuan Fund for Mathematics of National Natural Science Foundation of China (NSFC) (12326608), Shenzhen Key Laboratory of Cross-Modal Cognitive Computing (grant number ZDSYS20230626091302006), and Shenzhen Stability Science Program 2023.

## Limitations

While MLLM-Bench strives to assess multimodal large language models (MLLMs) comprehensively, it cannot encapsulate the full diversity of real-world multimodal interactions, acknowledging the challenge of simulating the unpredictable variety of real-life tasks.

**Potential Subjectivity** The design of human-annotated per-sample criteria, which seeks to mirror human user experience, may introduce subjectivity, potentially affecting the consistency and generalizability of results. However, the primary objective of introducing per-sample criteria is to align model performance with human needs, as the benchmark is designed to evaluate the real-world capabilities of the models. While we acknowledge the existence of subjectivity in per-sample criteria, we posit it as a "positive" subjectivity, as it brings the model’s performance closer to human expectations and practical needs.

**Reproducibility** To promote reproducibility, we have provided all the necessary code to replicate the results presented in this paper, along with the evaluation prompts detailed in Appendix E. Reproducibility is a cornerstone of this research, underscoring its importance and our commitment to transparency and scientific rigor.

**Potential Deficiency of GPT-4-Turbo** While GPT-4-Turbo is a potential evaluator, its assessment outcomes may not always align perfectly with factual accuracy or human standards. There are situations where GPT-4-Turbo itself cannot handle the task. If GPT-4-Turbo itself cannot solve the problem, it cannot be a qualified judge to conduct the evaluation. Table 12 shows examples of GPT-4-Turbo’s incorrect evaluation without our per-sample criteria.

**Potential Biases** The mention of length and position biases points to intrinsic limitations in GPT-4-Turbo’s processing, as discussed in Section 6.4. These biases can affect the model’s performance on the benchmark, potentially skewing results based on the length of input or the position of relevant information. This suggests a need for criteria that account for these biases, ensuring that the evaluation reflects the model’s ability to understand and generate content impartially, regardless of these factors. As mentioned in Section 5.1, shuffling the order of model-generated responses could migrate the position bias, as done by (Wang et al., 2023b; Chen et al., 2023b).

## Extensibility to Larger Scale Benchmark

**Dataset** One might be concerned that the current dataset size is not large enough and the dataset’s category distribution is unbalanced. In this paper, the selection of 420 samples was a deliberate methodological choice, aimed at demonstrating the efficacy of our proposed evaluation paradigm (i.e. per-sample criteria for evaluation) for multimodal large language models (MLLMs) in open-ended tasks. Moreover, our approach is scalable. One can scale up our benchmark through few-shot prompting, where new images are collected, and GPT-4-Turbo is used to generate instructions and corresponding per-sample criteria based on the samples and criteria in our benchmark. While there is a trade-off between scalability and the precision of per-sample criteria, one effective solution is to manually verify the generated criteria. This process strikes a balance between reducing time and human resource costs while maintaining a precise and comprehensive evaluation. Also, we are committed to continuously updating and improving our benchmark to ensure a more balanced and comprehensive system in future versions.

**Extensibility to More Recent MLLMs** Within the scope of our continuous endeavors, we remain dedicated to the inclusion of the most recent models into our evaluative framework. We have established an online leaderboard, which openly encourages submissions from the global community. This platform is designed to facilitate the perpetual evaluation and juxtaposition of novel models against pre-established benchmarks, thereby promoting an atmosphere of relentless innovation and enhancement. Notably, to prevent evaluation leakage and deter potential benchmark manipulation, the per-sample criteria employed by this online

leaderboard are confidential and accessible exclusively to individual submitters.

**Societal Impact** By promoting automated evaluation methods, there is a risk that human judgment may be undervalued or overlooked. While our benchmark aims to align closely with human evaluations, it is crucial to maintain a balance and ensure that human oversight remains integral to the evaluation process.

## Ethical Statements

There are no ethics-related issues in this paper. The data and other related resources in this work are open-source and commonly used by many existing studies.

## References

- Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. [Gemini: A family of highly capable multimodal models](#). *CoRR*, abs/2312.11805.
- Anthropic. 2024. [The claude 3 model family: Opus, sonnet, haiku](#).
- Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak 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](#). *CoRR*, abs/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](#).
- Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. 2023b. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*.
- Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Saĝnak 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. *arXiv preprint arXiv:2308.06595*.
- Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. [Humans or llms as the judge? a study on judgement biases](#).
- Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024b. [Allava: Harnessing gpt4v-synthesized data for A lite vision-language model](#). *CoRR*, abs/2402.11684.

- Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. [Minigpt-v2: large language model as a unified interface for vision-language multi-task learning](#).
- Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, et al. 2023b. Phoenix: Democratizing chatgpt across languages. *arXiv preprint arXiv:2304.10453*.
- 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](#).
- 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*.
- Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. [Blink: Multimodal large language models can see but not perceive](#).
- Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2023. [Making llama see and draw with seed tokenizer](#).
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*.
- David R Krathwohl. 2002. A revision of bloom’s taxonomy: An overview. *Theory into practice*, 41(4):212–218.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.
- Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. 2023b. [Empowering vision-language models to follow interleaved vision-language instructions](#). *CoRR*, abs/2308.04152.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. [Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#).
- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023d. AlpacaEval: An automatic evaluator of instruction-following models. [https://github.com/tatsu-lab/alpaca\\_eval](https://github.com/tatsu-lab/alpaca_eval).
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023e. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.
- Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. HallusionBench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*.
- Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024a. [World model on million-length video and language with ringattention](#). *CoRR*, abs/2402.08268.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. [Improved baselines with visual instruction tuning](#).
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. [Llava-next: Improved reasoning, ocr, and world knowledge](#).
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. 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. 2023d. MMBench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.
- OpenAI. 2023. [Gpt-4v\(ision\) system card](#).
- Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. [Kosmos-2: Grounding multimodal large language models to the world](#).
- Junke Wang, Lingchen Meng, ZeJia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023a. [To see is to believe: Prompting gpt-4v for better visual instruction tuning](#).
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023b. [Large language models are not fair evaluators](#).
- Minghao Wu and Alham Fikri Aji. 2023. [Style over substance: Evaluation biases for large language models](#).



- Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*.
- 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](#).
- Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.
- Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. 2023. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#).
- Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. 2022. [Detecting twenty-thousand classes using image-level supervision](#).
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

## A Detailed taxonomy

The detailed taxonomy is shown in Table 9. The table provides a comprehensive overview of 42 capabilities distributed across six cognitive levels within the MLLM-Bench framework. These capabilities are designed to benchmark the performance of multimodal language and vision models (MLLMs) across a range of tasks that mimic human cognitive abilities. Here’s a summary of the capabilities by level:

- **Level 1: Perception** - This level focuses on basic recognition tasks such as identifying objects, symbols, actions, and landmarks, as well as recognizing text in various languages and formats.
- **Level 2: Understanding** - At this level, capabilities extend to more complex comprehension tasks like scene and attribute recognition, understanding emotions and facial expressions, recognizing hidden objects, and grasping multimodal commonsense, including jokes and memes, across different languages and cultures.
- **Level 3: Applying** - This involves applying knowledge to practical tasks, including localizing objects, counting, understanding spatial relationships, interpreting medical images, and generating image captions and dense captions that describe images in detail.
- **Level 4: Analyzing** - Here, the focus shifts to analytical tasks such as understanding natural relations, analyzing charts, comparing attributes, finding differences, reasoning about events and social relationships, deducing identity and function, reasoning about physical properties, and predicting actions and trends.
- **Level 5: Evaluation** - This level assesses the model’s judgment capabilities, including evaluating image quality, assessing damage, detecting fake images, and identifying ethical issues.
- **Level 6: Creation** - The highest level of cognitive capability involves creative tasks, such as using vision to aid in coding and telling stories visually.

Each level builds on the previous ones, progressing from basic perception to complex and creative

problem-solving, reflecting an ascending order of cognitive complexity and capability required by MLLMs.

## B More Details on Data Collection and Annotation

We mainly introduce the division of labor for these six volunteers as well as the pipeline for data collection and annotation.

### B.1 Task Distribution for Volunteers

To ensure consistency and quality across our dataset, we have engaged six volunteers, each of whom is tasked with collecting and annotating data within one or two specific levels. The distribution of responsibilities is as follows:

- **Perception Level:** One volunteer is responsible for all 70 instances.
- **Understanding Level:** Two volunteers share this category, with each annotating 55 instances.
- **Applying Level:** A single volunteer manages all 60 instances.
- **Analyzing Level:** This is shared between two volunteers, with each handling 60 instances.
- **Evaluation and Creation Levels:** A single volunteer is responsible for the combined total of 60 instances across these two categories to balance the workload.

This division ensures that the workload is approximately equal for each volunteer. The entire annotation cycle was completed over a span of 21 days.

### B.2 Guideline

The guideline for data collectors in Table 10 outlines essential considerations for data annotation to ensure the dataset’s quality and relevance. Firstly, data currency is emphasized by prioritizing the most recent images, ideally within three months, sourced from social networks or daily life captures to mitigate the risk of data leakage. License-friendliness is crucial, advocating for the use of publicly licensed data with clear sharing agreements to respect intellectual property rights. Image clarity is necessary, with a recommendation for a mix of high and lower resolution images (with a

Capability Level	Capability
Level 1: Perception	General Object Recognition OCR Multilingual Text Recognition Action Recognition Symbol Recognition Food Recognition Landmark Recognition
Level 2: Understanding	Scene Understanding Attribute Recognition Image Topic Understanding Hidden Objects Recognition Facial Expression Recognition Emotion Understanding Multimodal Commonsense Understanding Joke and Meme Understanding Multilingual Multicultural Understanding Document Understanding Table Understanding
Level 3: Applying	Object Localization Object Counting Spatial Relationship Understanding Medical Image Understanding Image Captioning Dense Captioning
Level 4: Analyzing	Natural Relation Understanding Chart Understanding Attribute Comparison Difference Finding Event Cause Reasoning Social Relation Reasoning Identity Reasoning Function Reasoning Physical Property Reasoning Visual Math Reasoning Action Prediction Trend Prediction
Level 5: Evaluation	Image Quality Evaluation Damage Evaluation Fake Image Detection Ethical Problem Detection
Level 6: Creation	Coding Capability with Vision Visual Storytelling

Table 9: Overview of 42 capabilities on 6 cognitive levels in MLLM-Bench.

minimum short-edge resolution of 512) to accurately represent real-world conditions. Impartiality is maintained by avoiding content related to sensitive topics, ensuring the dataset’s neutrality and broad applicability. Instruction-image cohesion is vital, requiring precise and contextually reflective instructions to facilitate clear understanding and effective model testing. Lastly, diverse response formats are encouraged to enrich the dataset, promoting varied feedback beyond simple binary choices to better simulate real-world interactions and enhance analytical capabilities.

## C Data Statistics

MLLM-Bench is characterized by a rich diversity and complexity of instructions, each tailored to probe a specific capability of multimodal large language models. These instructions challenge models to generate responses that are both comprehensive and descriptive, engaging with the multifaceted nature of real-world scenarios and information. To illustrate the breadth of our instruction set, we present a word cloud visualization that encapsulates the frequency of terms within our instructions, as shown in Figure 4. We list one example per category in Table 4.

Characteristics	Description
Data Currency	To mitigate the risk of data leakage, it is crucial to utilize the most current images available. Annotators should <b>prioritize images sourced from social networks like Twitter or directly captured in daily life</b> , ideally within three months prior to our data collection phase. This approach helps prevent the collected data from being previously used in the training of evaluated models. Maintaining and updating the dataset in real time is recommended to address potential data leakage concerns effectively.
License-friendliness	When selecting data sources, prioritize those that are <b>publicly licensed and offer favorable sharing agreements</b> . Social networks like Twitter or content personally photographed by annotators are preferred due to their clear copyright status. This ensures that our data collection respects intellectual property rights and adheres to legal requirements.
Image Clarity	The collected images must be of sufficient quality to be identifiable by humans. However, incorporating a selection of lower-resolution images ( <b>the short-edge resolution is at least 512</b> ) is also advised to accurately represent the diversity of data quality encountered in real-world scenarios. This approach ensures that our dataset reflects practical conditions and challenges.
Impartiality	To maintain the dataset’s neutrality, <b>avoid including content related to sensitive topics</b> such as geopolitical issues. This commitment to impartiality ensures that our dataset can be used widely without bias, supporting a broad range of applications and studies.
Instruction-Image Cohesion	To ensure images and instructions are related, the instructions should not only be precise, but also be tailored to reflect the context depicted in the image as well as to test the specific model capability. For this to be effective, the language used in the instructions should be precise and unambiguous, facilitating a clear understanding of the image content and task requirements.
Diverse Responses Formats	Encouraging a variety of response formats enriches the dataset and better simulates real-world interactions. Instead of limiting responses to simple binary choices, data collectors are encouraged to seek out descriptive narratives and diverse forms of feedback. This strategy distinguishes our dataset from existing benchmarks and enhances its applicability to complex analytical tasks.

Table 10: The guideline for data collectors that states key considerations for data annotation.

## D Details of Benchmarked Models

Table 11 presents a comparison of various models, highlighting their characteristics such as their open-source availability and architectural components, including visual adapters and base large language models (LLMs). Models vary in size from 1.7B to 14B parameters, with some details remaining confidential, indicated by "/". Not all models are open-sourced, as exemplified by GPT-4-Turbo. Architectural details vary, with some models utilizing visual adapters like CLIP-ViT-L, ViT+Q-Former, and others, while the base LLMs mentioned include Vicuna-13B, LLaMA2-7B, and more. The table aims for clarity by abbreviating model names and provides a snapshot of the diverse approaches in integrating visual processing with language models, demonstrating a range of strategies for enhancing model capabilities.



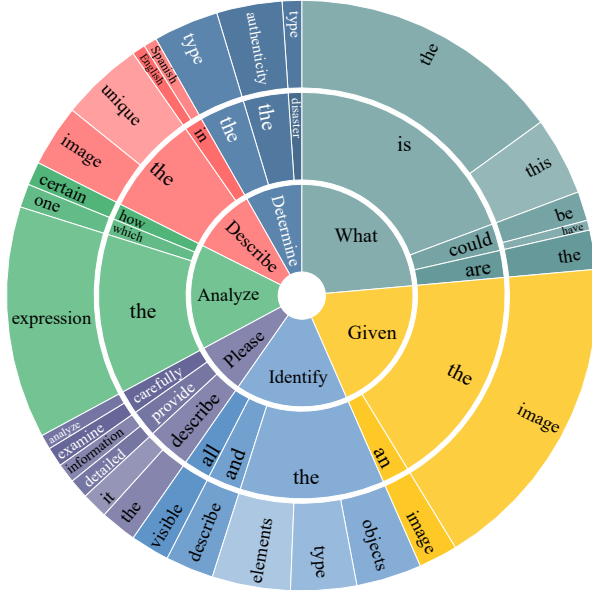


Figure 4: Distribution of questions in MLLM-Bench. We present the relative distribution of these recurring Instructions and their subsequent distributions.

## E Details of GPT-4-Turbo Evaluation

### E.1 Potential Deficiency of GPT-4-Turbo

While GPT-4-Turbo is a potential evaluator, its assessment outcomes may not always align perfectly with factual accuracy or human standards. There are situations where GPT-4-Turbo itself cannot handle the task. If GPT-4-Turbo itself cannot solve the problem, it cannot be a qualified judge to conduct the evaluation. Table 12 shows examples of GPT-4-Turbo’s incorrect evaluation without our per-sample criteria.

### E.2 Prompt for Pairwise Voting Using GPT-4V with Per-Sample Criteria

Our evaluations are conducted following a *pairwise voting* protocol. A stronger model is expected to have a larger number of wins in pairwise voting. The prompt of conducting voting by GPT-4-Turbo is shown in Figure 6.

### E.3 Anchors in Pairwise Evaluation

To facilitate a fair and consistent comparison across multiple models, we employ an anchor-based evaluation strategy. For each protocol, we use answers from the moderately powerful model *LLaVA-v1.5-Vicuna-13B* as benchmarks.

## F Additional Results for Validation Experiments

### F.1 Correlation with Human Alignment

**Model Selection** Our test set for the experiment is constructed by extracting two data points from each of the 42 capabilities identified in our benchmark. We select three models representing different levels of capabilities—top, middle, and bottom—as determined by GPT-4-Turbo’s direct voting outcomes. These models are **Qwen-VL-Chat**, **InstructBLIP-Vicuna 13B**, and **BLIP2-Flan-T5-XL**. **LLaVA-v1.5-13B** serves as an anchor. The evaluation set thus comprised 252 items in total. All volunteers are required to make judgments on all 252 data samples.

**Interface for Human Evaluation** The user interface of human evaluation is shown in Figure 5.

**Detailed Results** Table 15 presents the pairwise agreement among four human evaluators, offering insights into inter-evaluator alignment levels regarding a set of evaluations. The inter-evaluator agreement is calculated by the matching percentage of two evaluators’ voting results. The agreement ranges from 0.80 to 0.88, indicating a high degree of consensus among the evaluators. Specifically, the correlation values suggest that while there is a strong overall alignment in their assessments, each evaluator also brings a unique perspective to the evaluation process. Evaluator 1 and Evaluator 2 exhibit the highest correlation (0.88), suggesting their evaluations are most closely aligned. In contrast, the lowest correlation is observed between Evaluator 1 and Evaluator 3 (0.69), indicating a lesser, yet still significant, level of agreement. These findings underscore the evaluators’ ability to consistently recognize and rank the evaluated items according to similar criteria, while also maintaining individual discretion in their judgments. This balance between consensus and individuality is crucial for ensuring both the reliability and the richness of the evaluation process, highlighting the evaluators’ competence in providing nuanced assessments.

### F.2 Bounding Box Generation

We employ Detic<sup>7</sup> (Zhou et al., 2022) to generate bounding box coordinates and their labels. For each image, we only keep the top-10 objects according

<sup>7</sup>We use the version Detic\_LCOCO121k\_CLIP\_SwinB\_896b32\_4x\_ft4x\_max-size.

Models	MLLM Architecture	GitHub Stars	Huggingface Download
<i>Closed-Source</i>			
GPT-4-Turbo (OpenAI, 2023)	/		
Claude-3 (Anthropic, 2024)	/		
Gemini-Pro (Anil et al., 2023)	/		
<i>Open-Source</i>			
LLaVA-v1.5-13B (Liu et al., 2023b)	Pretrained Vision Encoder + Projector + LLM	15.4K	333.7K
LVIS-Instruct4v-LLaVA-7B (Wang et al., 2023a)	Pretrained Vision Encoder + Projector + LLM	122	5
MiniGPT-v2 (Chen et al., 2023a)	Pretrained Vision Encoder + Projector + LLM	24.7K	/
LLaVA-v1.5-7B (Liu et al., 2023b)	Pretrained Vision Encoder + Projector + LLM	15.4K	703K
LLaVA-v1.6-Vicuna-7B (Liu et al., 2024b)	Pretrained Vision Encoder + Projector + LLM	15.4K	1.2M
LLaVA-v1.6-Vicuna-13B (Liu et al., 2024b)	Pretrained Vision Encoder + Projector + LLM	15.4K	100.1K
LLaVA-v1.6-34B (Liu et al., 2024b)	Pretrained Vision Encoder + Projector + LLM	15.4K	592.8K
Yi-VL-6B (Young et al., 2024)	Pretrained Vision Encoder + Projector + LLM	7K	17.2K
ALLaVA-3B-Longer (Chen et al., 2024b)	Pretrained Vision Encoder + Projector + LLM	134	93
kosmos2 (Peng et al., 2023)	Pretrained Vision Encoder + Grounded LLM	18.1K	29.2K
LWM (Liu et al., 2024a)	Pretrained Vision Encoder + Projector + Long-Context LLM	6.6K	/
BLIP2-Flan-T5-XL (Li et al., 2023c)	Query tokens + LM	8.5K	35.4K
Qwen-VL-Chat (Bai et al., 2023a)	Query tokens + LLM	3.4K	289.9K
InstructBLIP-Vicuna-13B (Dai et al., 2023)	Query tokens + LLM	8.5K	5.4K
mPLUG-Owl2 (Ye et al., 2023)	Query tokens + LLM with Modality-Adaptive Module	1.9K	9.7K
Cheetor (Li et al., 2023b)	Query tokens + VPG-C + LLM	308	/
Fuyu-8B (Bavishi et al., 2023)	Linear Vision Encoder + LLM	/	17.9K
SEED-LLaMA (Ge et al., 2023)	VQ-based Vision Encoder + LLM	445	/
OpenFlamingo (Awadalla et al., 2023)	Perceiver Resampler + LLM with Gated Cross-Attention Layers	3.4K	/

Table 11: Model architecture and popularity. "/" means the model either uses a private download link or their download counts on HuggingFace are not shown.

Question

9. Given the image, assess the floor plan's design in terms of privacy for the bedrooms and accessibility to the communal areas.

Answers

The floor plan of the bedrooms features a two bedrooms and a living room, with a hallway connecting them. Additionally, there is a shared bathroom between the two bedrooms, providing privacy for each individual. Furthermore, the living room shares an open space with the bedrooms, ensuring accessible access for all occupants. This design can be seen as a way to create a compact, but comfortable living space for the three people.

The floor plan, which features bedrooms, a dining area, and a living area, appears to have a well-thought-out design in terms of privacy and accessibility. The bedrooms are located on the upper floor, while the dining and living areas are on the main floor. This arrangement helps separate the sleeping areas from the communal spaces and provides some privacy for the occupants of the bedrooms. Additionally, the layout designates certain spaces for specific purposes, such as the dining area and the living area, which ensures a clear distinction between these zones. This further contributes to the accessibility and ease of movement within the house. Moreover, the floor plan shows a clear layout for the kitchen, with a sink and a refrigerator positioned in strategic locations, which suggests the kitchen's design is efficient and functional. Overall, the floor plan seems to consider the privacy and accessibility needs of its users, facilitating a comfortable and functional living environment.

☐ Unable to decide

Criteria

1. Evaluate if bedrooms are isolated from communal areas to ensure privacy.

2. Assess ease of movement from bedrooms to communal spaces like kitchen and living.

Confirm

Image

SIMPLE FLOOR PLAN

Figure 5: User interface of human evaluation.

**Question:** Determine the percentage increase in 'Total Income' from the quarter 30 Sep 2023 to the quarter 30 June 2023.

**YATHARTH HOSPITAL & TRAUMA CARE SERVICES LIMITED**

Regd. Office : JA 108 DLF Tower A NA Jasola District Centre South Delhi DL 110025 India

Corporate Office: HO-01, Sector-1 Greater Noida West Gautam Buddha Nagar UP 201306 India

Website: www.yatharthhospitals.com

Email: cs@yatharthhospitals.com

CIN : U85110DL2008PLC174706

STATEMENT OF UNAUDITED CONSOLIDATED FINANCIAL RESULTS FOR THE QUARTER/ HALF YEAR ENDED 30TH SEPTEMBER, 2023

Particulars	CONSOLIDATED					
	Quarter Ended			Half Year Ended		Year Ended (Audited)
	30 Sep'2023	30 June'2023	30 Sep'2022	30 Sep'2023	30 Sep'2022	31 March'2023
I Revenue from operations	1,712.80	1,545.07	1,278.89	3,257.86	2,389.56	5,202.93
II Other income	38.21	7.57	7.45	45.77	13.30	28.07
III Total Income (I+II)	1,751.00	1,552.63	1,286.34	3,303.64	2,402.86	5,231.00
IV Expenses						
Medical consumables and pharmacy items consumed	360.47	280.52	230.99	640.98	443.91	929.35
Employee benefits expense	283.93	268.38	237.35	552.31	432.70	919.30
Finance costs	29.41	59.29	49.80	88.71	94.76	213.87
Depreciation and amortisation	68.79	68.74	62.79	137.53	123.48	275.07
Other expenses	612.70	582.02	475.32	1,194.73	920.89	2,016.63
Total expenses (IV)	1,355.30	1,258.95	1,056.25	2,614.25	2,015.74	4,354.22

**Answer 1 (LLaVA-v1.5-13B):**

To calculate the percentage increase in 'Total Income' from the quarter ended 30 Sep 2023 to the quarter ended 30 June 2023, we need to use the formula: (Current value - Previous value) / Previous value x 100. The Total Income values for the two quarters are as follows: 30 Sep 2023: 1,659,059 30 June 2023: 1,174,573 Using the formula, the percentage increase in Total Income is calculated as follows: (1,659,059 - 1,174,573) / 1,174,573 x 100 = 43.816% So, there is a 43.816% increase in Total Income from the quarter ended 30 Sep 2023 to the quarter ended 30 June 2023.

**Answer 2 (Fuyu-8B):**

240.28

- Criteria:** 1. Correctly calculates percentage increase from 'Total Income' for specified quarters.  
2. Matches the percentage increase with the reference: **12.78%**.  
3. Explains the calculation process or provides the correct formula used.

**GPT-4-Turbo Judgment (w/o criteria)**

**GPT-4-Turbo Judgment (w/ criteria)**

Answer 1

Unable to decide: situation one

Table 12: Examples showing GPT-4-Turbo's capability fall short.

**Prompt for Pairwise Voting using GPT-4-Turbo**

**Prompt:**

### You are an excellent evaluator.

### Your assignment involves providing evaluations for given responses.

### Each evaluation consists of \*an image\*, \*a question\*, a \*question type\*, and \*two corresponding answers\*. Your task is to discern which answer is superior based on the \*\*quality\*\* and its alignment w.r.t the image.

### There are only two situations where you may choose 'unable to decide':

#### Situation one: The question type is 'close-ended' and both answers are correct or wrong.

#### Situation two: Both answers contain considerable factual errors or ethical issues.

### Otherwise, you should always choose a better answer by responding 'Answer1' or 'Answer2'.

### You should ONLY output your vote 'Answer1', 'Answer2', 'unable to decide: situation one', or 'unable to decide: situation two' in the last line.

~~~Question

{question}

~~~

~~~Question Type

{question\_type}

~~~

~~~Answer1

{answer1}

~~~

~~~Answer2

{answer2}

~~~

### Please refer to the given criteria when you making the judgment

Criteria: {criteria}

Figure 6: The prompt used for Directing Voting using GPT-4-Turbo.

to confidence scores. For each object, we store the coordinates of top-left and bottom-right corner, which will later be sent to text-only GPT-4.

### F.3 Prompts for Caption Generation

The prompt for generating image captions is shown in Figure 7.

### F.4 Prompts for GPT-4 Evaluation

We show the evaluation prompts for *GPT-4 + caption* and *GPT-4 + detection* in Figure 8 and Figure 9, respectively.

### F.5 Prompt perturbation

We supplement two settings varying the prompts to investigate the impact of the formulation of prompts on our evaluation results. In the first setting, we aim to assess how variations in prompt content could affect GPT-4-Turbo evaluation, we reformed the prompts (change wordings and prompt structure, while the content remains semantically the same) by ‘gpt-4-0125-preview’. The rewritten prompt is shown in Figure 10. In the second setting, we incorporate an extra requirement in the prompt, specifically instructing the judge to conduct a Chain of Thought (CoT) after making the selection. The reformed prompt is shown in Figure 11. Results for different settings are summarized in Table 13.

### F.6 Comparison of Alignment between Human and Different Methods

Table 14 shows the comparison of alignment between different evaluation methods and human evaluation. The number of evaluations for each settings are listed below the ratio.

## G Correlation with Other Benchmarks

We discuss the correlation between MME-Perception, MME-Cognition and MM-Vet and MLLM-Bench. Specifically, we compute the pairwise Spearman correlation of rankings. Table 16 shows that MLLM-Bench has the highest correlation with MM-Vet, which is a comprehensive benchmark testing 6 core abilities of MLLMs with short-answers. On the other hand, the Perception and Cognition splits of MME evaluate MLLMs at a fundamental and an advanced level, respectively. Therefore, the latter split has a higher correlation with MLLM-Bench than the former does. Thus, we highlight that our benchmark is proposed to

complement with previous works, testing MLLMs at multiple levels with open-ended questions.



#### Prompt for generating image captions using GPT-4-Turbo

##### Prompt:

```

### You are an excellent image describer and captioner
### Your task is to describe the given image as detailed as possible and give a caption for the image
### Do not mention anything from the prompt in your response
### You will follow the instructions to the best of your ability
### Your response should follow the following format
<start of description>
{description}
<end of description>

```

Figure 7: Prompts for generating image captions using GPT-4-Turbo.

#### Prompts for *GPT-4 + Caption Evaluation*

##### Prompt:

```

### You are an excellent evaluator.
### Your assignment involves providing evaluations for given responses.
### Each evaluation consists of *a caption*, *a question*, a *question type*, and *two corresponding answers*. Your task is
to discern which answer is superior based on the **quality** and its alignment w.r.t the caption.
### There are only two situations where you may choose 'unable to decide':
#### Situation one: The question type is 'close-ended' and both answers are correct or wrong.
#### Situation two: Both answers contain considerable factual errors or ethical issues.
### Otherwise, you should always choose a better answer by responding 'Answer1' or 'Answer2'.
~~~Caption
{caption}
~~~
~~~Question
{question}
~~~
~~~Answer1
{answer1}
~~~
~~~Answer2
{answer2}
~~~
### You should ONLY output your vote 'Answer1', 'Answer2', 'unable to decide: situation one', or 'unable to decide:
situation two' in the last line.

```

Figure 8: Prompts for *GPT-4 + Caption Evaluation*

Settings	Perception	Understanding	Applying	Analyzing	Evaluation	Creation	Total
Default	0.49	0.35	0.43	0.35	0.38	0.23	0.37
Setting1: Prompt Reforming	0.50	0.34	0.32	0.36	0.30	0.28	0.36
Setting2: Answer-then-CoT	0.49	0.36	0.42	0.33	0.33	0.35	0.38

Table 13: Prompt robustness analysis for Qwen-VL-Chat using LLaVA-v1.5-13B as anchor. All values represent win rates.

### Prompts for GPT-4 + Detection Evaluation

#### Prompt:

### You are an excellent evaluator.  
 ### Your assignment involves providing evaluations for given responses.  
 ### Each question will consist of a \*list of objects\* about an image, a \*question\* and \*two corresponding answers\*. Your task is to discern which response is superior based on the \*\*quality of the answer\*\* and its alignment w.r.t the objects.  
 ### Each object in the object list will contains three keys, "bbox", "conf", "label"  
 ##### "bbox" is a list of four numbers, which are the coordinates of the bounding box of the object in the image, the order is [x1, y1, x2, y2], where (x1, y1) is the top left corner of the bounding box, (x2, y2) is the bottom right corner of the bounding box.  
 ##### "conf" is a number, which is the confidence of the object detection model.  
 ##### "label" is a string, which is the label of the object.  
 ### There are only two situations where you may choose 'unable to decide':  
 ##### Situation one: The question type is 'close-ended' and both answers are correct or wrong.  
 ##### Situation two: Both answers contain considerable factual errors or ethical issues.  
 ### Otherwise, you should always choose a better answer by responding 'Answer1' or 'Answer2'.

~~~Object List

{object\_list}

~~~

~~~Question

{question}

~~~

~~~Answer1

{answer1}

~~~

~~~Answer2

{answer2}

~~~

### You should ONLY output your vote 'Answer1', 'Answer2', 'unable to decide: situation one', or 'unable to decide: situation two' in the last line.

Figure 9: Prompts for GPT-4 + Detection Evaluation

	Evaluator 1 (Order 1)	Evaluator 2 (Order 1)	Evaluator 3 (Order 2)	Evaluator 4 (Order 2)	Aggregated
GPT-4 + detection w/o Criteria	79.1% (106/134)	78.2% (104/133)	71.54% (93/130)	72.48% (79/109)	75.37% (101/134)
GPT-4 + detection w/ Criteria	82.71% (110/133)	78.79% (104/132)	72.87% (94/129)	80.18% (89/111)	81.2% (108/133)
GPT-4 + caption w/o Criteria	80.0% (152/190)	81.72% (152/186)	74.59% (135/181)	82.14% (138/168)	81.15% (155/191)
GPT-4 + caption w/ Criteria	78.72% (148/188)	80.32% (151/188)	74.58% (132/177)	80.98% (132/163)	80.53% (153/190)
GPT-4 + detection + caption w/o Criteria	81.12% (159/196)	83.42% (161/193)	75.27% (140/186)	84.12% (143/170)	83.0% (166/200)
GPT-4 + detection + caption w/ Criteria	80.41% (156/194)	82.29% (158/192)	74.86% (137/183)	83.54% (137/164)	83.16% (163/196)
GPT-4-Turbo w/o Criteria	81.62% (151/185)	82.22% (148/180)	75.42% (135/179)	82.05% (128/156)	82.8% (154/186)
GPT-4-Turbo w/ Criteria	<b>78.61%</b> (158/201)	<b>80.30%</b> (159/198)	<b>73.00%</b> (146/200)	<b>77.84%</b> (144/185)	<b>88.02%</b> (147/167)

Table 14: Comparison of alignment between different evaluation methods and human evaluation, the aggregated column aggregates the results of 4 human evaluators (includes the number of valid evaluations in each setting). The agreement is calculated through a matching percentage.

#### Rewritten prompt for pairwise voting using GPT-4-Turbo

##### **Prompt:**

### You possess exceptional assessment skills.  
### Your task is to conduct evaluations on provided responses.  
### Each evaluation will include a visual, a query, a category of query, and two possible responses. Your role is to determine which response is of higher quality and how well it aligns with the visual.  
### You are permitted to select 'unable to decide' only under two specific situations:  
#### Situation one: The query falls under the 'close-ended' category and both responses are either correct or incorrect.  
#### Situation two: Both responses exhibit significant factual inaccuracies or ethical concerns.  
### In all other situations, you are obligated to identify a more suitable response by indicating 'Answer1' or 'Answer2'.  
### Your decision should be only stated as 'Answer1', 'Answer2', 'unable to decide: situation one', or 'unable to decide: situation two' in your final response.

~~~Question

{question}

~~~

~~~Question Type

{question\_type}

~~~

~~~Answer1

{answer1}

~~~

~~~Answer2

{answer2}

~~~

### Please refer to the given criteria when you making the judgment

Criteria: {criteria}

Figure 10: Rewritten prompt for pairwise voting using GPT-4-Turbo.

#### Prompt with CoT for pairwise voting using GPT-4-Turbo

##### **Prompt:**

### You are an excellent evaluator.  
### Your assignment involves providing evaluations for given responses.  
### Each evaluation consists of an image, a question, a question type, and two corresponding answers. Your task is to discern which answer is superior based on the quality and its alignment w.r.t the image.  
### There are only two situations where you may choose 'unable to decide':  
#### Situation one: The question type is 'close-ended' and both answers are correct or wrong.  
#### Situation two: Both answers contain considerable factual errors or ethical issues.  
### Otherwise, you should always choose a better answer by responding 'Answer1' or 'Answer2'.  
### You should ONLY output your vote 'Answer1', 'Answer2', 'unable to decide: situation one', or 'unable to decide: situation two' in the first line, and then give out your reason.

~~~Question

{question}

~~~

~~~Question Type

{question\_type}

~~~

~~~Answer1

{answer1}

~~~

~~~Answer2

{answer2}

~~~

### Please refer to the given criteria when you making the judgment

Criteria: {criteria}

Figure 11: Prompt with CoT for pairwise voting using GPT-4-Turbo.

	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4
Evaluator 1	-	0.88	0.80	0.82
	-	(198/225)	(177/220)	(160/196)
Evaluator 2	0.88	-	0.85	0.83
	(198/225)	-	(181/212)	(161/194)
Evaluator 3	0.80	0.85	-	0.81
	(177/220)	(181/212)	-	(153/189)
Evaluator 4	0.82	0.83	0.81	-
	(160/196)	(161/194)	(153/189)	-

Table 15: Pairwise Agreement among Human Evaluators.

	$\text{MME}^P$	$\text{MME}^C$	MM-Vet	<b>MLLM-Bench</b>
$\text{MME}^P$	1.00	0.14	-0.40	0.57
$\text{MME}^C$	0.14	1.00	0.80	0.81
MM-Vet	-0.40	0.80	1.00	0.89
<b>MLLM-Bench</b>	0.57	0.81	0.89	1.00

Table 16: Pairwise Spearman correlation between different benchmarks:  $\text{MME}^P$ : MME-Perception;  $\text{MME}^C$ : MME-Cognition; MM-Vet; **MLLM-Bench**.