

MMEVALPRO: Calibrating Multimodal Benchmarks Towards Trustworthy and Efficient Evaluation

Jinsheng Huang^{1,2,3*}, Liang Chen^{1,2*}, Taian Guo^{1,2,3}, Fu Zeng⁴, Yusheng Zhao^{1,2,3}
 Bohan Wu^{1,2,3}, Ye Yuan^{1,2,3}, Haozhe Zhao¹, Zhihui Guo⁵, Yichi Zhang¹, Jingyang Yuan^{1,2,3}
 Wei Ju^{1,2,3}, Luchen Liu^{1,2,3}, Tianyu Liu⁶, Baobao Chang^{1,2†}, Ming Zhang^{1,2,3†}

¹National Key Laboratory for Multimedia Information Processing, Peking University

²School of Computer Science, Peking University, ³PKU-Anker LLM Lab

⁴Chinese Academy of Medical Sciences, ⁵CUHK, ⁶Alibaba Group

hjs@stu.pku.edu.cn leo.liang.chen@outlook.com {chbb,mzhang_cs}@pku.edu.cn

<https://mmevalpro.github.io>

Abstract

Large Multimodal Models (LMMs) exhibit impressive cross-modal understanding and reasoning abilities, often assessed through multiple-choice questions (MCQs) that include an image, a question, and several options. However, many benchmarks used for such evaluations suffer from systematic biases. Remarkably, Large Language Models (LLMs) without any visual perception capabilities achieve non-trivial performance, undermining the credibility of these evaluations. To address this issue while maintaining the efficiency of MCQ evaluations, we propose MMEVALPRO, a benchmark designed to avoid Type-I errors through a trilogy evaluation pipeline and more rigorous metrics. For each original question from existing benchmarks, human annotators augment it by creating one perception question and one knowledge anchor question through a meticulous annotation process. MMEVALPRO comprises 2,138 question triplets, totaling 6,414 distinct questions. Two-thirds of these questions are manually labeled by human experts, while the rest are sourced from existing benchmarks (MMMU, ScienceQA, and MathVista). Compared with the existing benchmarks, our experiments with the latest LLMs and LMMs demonstrate that MMEVALPRO is **more challenging** (the best LMM lags behind human performance by 31.73%, compared to an average gap of 8.03% in previous benchmarks) and **more trustworthy** (the best LLM trails the best LMM by 23.09%, whereas the gap for previous benchmarks is just 14.64%). Our in-depth analysis explains the reason for the large performance gap and justifies the trustworthiness of evaluation, underscoring its significant potential for advancing future research.

1 Introduction

Ever since the birth of standardized testing, the credibility of its conclusions has been a significant concern.

* Equal contribution. † Corresponding authors.

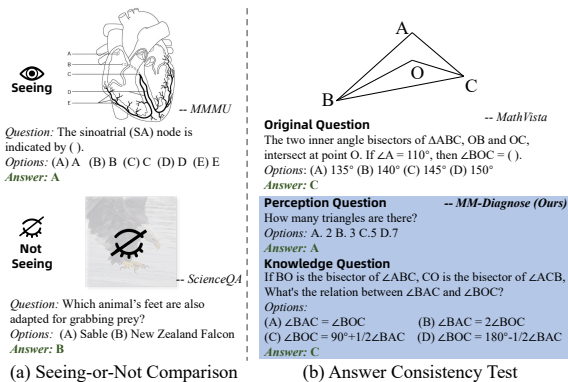


Figure 1: Examples of the probing experiments.

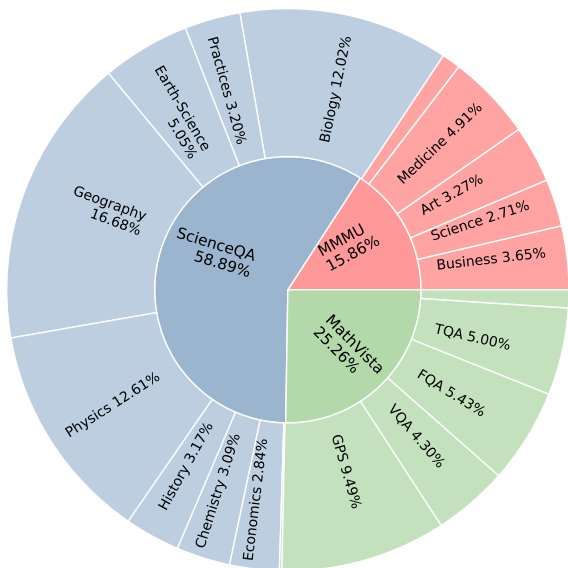


Figure 2: Topic distribution of MMEVALPRO's data.

The same problem goes for the evaluation of recently popular Large Multimodal Models (LMMs) such as GPT4-o (OpenAI, 2024), Gemini-1.5 (Team et al., 2024), Qwen-VL (Bai et al., 2023b) and LLaVA (Liu et al., 2023b). One classic composition of such an evaluation is the multiple-choice question (MCQ), which includes an image, a question, possible choices, and an answer. This form of evaluation has higher usability

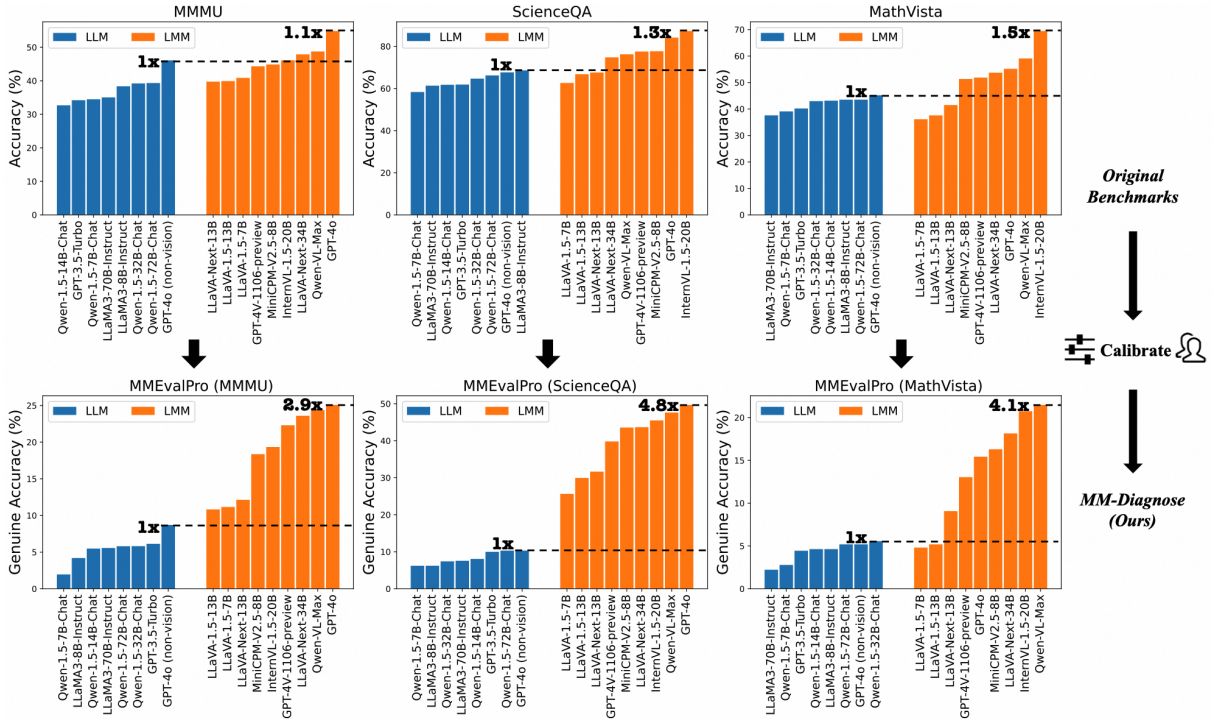


Figure 3: LLMs and LMMs’ performance comparison between original multimodal benchmarks and MMEVALPRO. Performance gap between LLM and LMM is much clearer in MMEVALPRO.

compared to other evaluation methods, such as text-based evaluation and human evaluation (Chen et al., 2024a). Many benchmarks (Fu et al., 2023; Liu et al., 2023c; Lu et al., 2024, 2022; Yue et al., 2023; Zhang et al., 2024a; Chen et al., 2023a, 2024b; Kembhavi et al., 2016; Shen et al., 2024) designed for multimodal foundation models include a large portion of MCQs and are widely adopted in testing from basic to most advanced multimodal models (OpenAI, 2024; Bai et al., 2023b; Liu et al., 2023b), with some models reaching or even surpassing human scores on certain benchmarks (Lu et al., 2022; Kembhavi et al., 2016). However, it is important to consider whether these evaluation results truly reflect the absolute capabilities of the models, especially when we are comparing them with human beings in the pursuit of AGI. In fact, several work such as PCA-Bench (Chen et al., 2024b), MMStar (Chen et al., 2024d) and MathVerse (Zhang et al., 2024b) have pointed out that the multimodal MCQ evaluation has intrinsic bias, which provides Large Language Models (LLMs) with shortcut to hack the question. FastV (Chen et al., 2024c) finds that LMM could even achieve better performance on some benchmarks with only partial visual tokens. In this paper, we mainly study three popular multimodal benchmarks MMMU (Yue et al., 2023), ScienceQA (Lu

et al., 2022) and MathVista (Lu et al., 2024).

In our preliminary Seeing-or-Not Comparison experiment, we found that LLMs could achieve high scores without processing the visual data, attributed to possible data leakage, visual information problems being not related to answering the question or simply guessing the answer. Notably, the average performance gap between best LLM and best LMM is just 14.64%, which is even smaller than the gap within LMM themselves, revealing the unreliable problem of such evaluations.

Our further Answer Consistency Test into multiple choice questions (MCQ) reveals a prevalent Type-I Error in such evaluation’s conclusion, where models could output correct answers without actual comprehension. For example, the model could calculate the degree for a particular angle, but could not recognize the correct angle’s name in the figure, which is a prerequisite to compute the degree.

To this end, we propose MMEVALPRO to truthfully reflect the true multimodal capabilities of tested models and keep the simplicity of MCQ evaluation. We achieve this by augmenting the original MCQ with prerequisite perception and knowledge questions. We propose Genuine Accuracy as the main metric, which depends on whether the model answers the triplet questions concurrently. Overall, in MMEVALPRO we annotate 2, 138 question

triplets, originating from MMMU, ScienceQA, and MathVista, resulting in 6,414 individual questions.

We carry out experiments and analyses involving 17 different models and human experts. The findings indicate that MMEVALPRO offers a more precise reflection of the tested LMMs’ capabilities and poses a more demanding challenge. MMEVALPRO is more trustworthy than base benchmarks as the best LLM trails the best LMM by 23.09% whereas the gap for previous benchmarks is just 14.64%. Significantly, even the most advanced models, including GPT-4o and Qwen-VL-Max, lag considerably behind human performance, with a notable gap of over 30% in Genuine Accuracy (only 8.03% for the base benchmarks). Our investigation into the factors contributing to the consistency gap illuminates the existing disparities and supports the evaluation credibility of MMEVALPRO, supplying insights for future research endeavors.

2 Probing the Credibility of Multimodal Benchmarks

The fundamental assumption of any benchmark is that models achieving higher scores possess superior capabilities. In this section, we question the credibility of such an assumption for existing multimodal benchmarks. We find that the existing benchmarks are not trustworthy enough in either relative or absolute perspectives, which is concluded from two probing experiments: Seeing-or-Not Comparison and Answer Consistency Test. The processes are illustrated in Figure 1. We test the MCQ evaluation in three multimodal benchmarks across various domains including MMMU, ScienceQA-Image and MathVista. We provide the detail of dataset statistics in section 3.1, model inference hyperparameters and prompts in Appendix-B.

2.1 Seeing-or-Not Comparison

As shown in Figure 1-(a), we prepare two data versions for each benchmark: “Seeing” (with image, question, options, and answer) and “Not-Seeing” (without image). We test leading LMMs on “Seeing” data and non-vision LLMs on “Not-Seeing” data, then compare the results. The outcomes are shown in the first row of Figure 3.

The figure indicates that the performance gap between LMMs and LLMs is significantly narrower than anticipated. Intuitively, one might assume that LLMs, which is unable to process visual information, would perform considerably worse

Dataset	Proportion
MMMU	2.97%
ScienceQA	43.08%
MathVista	5.37%

Table 1: Proportions of “Image is not needed” samples in the original datasets

on multimodal benchmarks. In fact, if we compare the scores between the best-performing LLM and LMM, we observe that for MMMU, the best LMM’s performance is only 1.1 times that of the best LLM. This performance gap is even smaller than the variability observed within LMMs. Similar results go for the other tested benchmarks. It’s more surprising that LLMs sometimes outperform their vision-enabled counterparts (GPT4-o without vision ability outperforms the LLaVA-1.5 series according to Figure 3) and there is not an apparent performance boundary between the two kinds of models. These results suggest that those benchmarks do not accurately reflect the true multimodal understanding capabilities of the tested models. The reason for this phenomenon is three-fold:

1. Image is not needed: Some benchmark questions can be answered solely through textual information, making visual input unnecessary. This diminishes the advantage of vision-enabled models (LMMs). For instance, as shown in the ScienceQA question in Figure 1, the knowledge that a falcon uses its feet to capture prey is common and does not require an image for verification.

2. Data leakage: During training, LLMs may inadvertently encounter similar questions or datasets, leading to unfair advantages. The existing benchmarks often derive questions from textbooks (Lu et al., 2022; Yue et al., 2023), online education resources, and research papers (Lu et al., 2024), which are also sources for training datasets (Touvron et al., 2023; Clement et al., 2019).

3. Educated guessing: LLMs are trained on extensive text datasets, enabling them to make educated guesses even without visual information, which narrows the performance gap with LMMs.

Quantitative Analysis We analysed the proportion of samples that can be answered correctly due to “Image is not needed” in the original datasets and the results are shown in the Table 1. As the results shown, the 43.08% proportion in ScienceQA is the highest, which is also reflected in the “Not-Seeing” experiments. From analysis, we can conclude that “Image is not needed” samples significantly

affected the credibility of previous benchmarks.

These factors collectively result in the unexpectedly narrow performance gap between LLMs and LMMs on multimodal benchmarks. To create a fair multimodal evaluation benchmark: (1) Ensure questions are intrinsically tied to the image details, making visual information essential for deriving the answer. (2) Prevent contamination of training data to ensure models are reasoning through problems instead of recalling memorized answers. (3) Design questions and answers to minimize the likelihood of the model making accurate guesses.

2.2 Answer Consistency Test

To determine whether a model truly understands a question or is simply "hacking" the answer, we simulate the human problem-solving process by creating "anchor questions" that must be answered before the main question. When humans tackle multimodal reasoning questions, they typically follow two key steps: (1) identifying relevant visual clues in the image, and (2) applying their knowledge to reason through the problem before arriving at the final answer. Omitting either step usually leads to an incorrect answer. Similarly, we create a perception anchor question and a knowledge anchor question related to the original question. If a model can answer both anchor questions and the final question, it demonstrates genuine comprehension and reasoning, rather than mere guessing.

An example is shown in Figure 1-(b), we set a perception question and a knowledge question to the MCQ from MathVista. For human examinees, the perception question and knowledge question are easier to answer than the original since they are the prerequisites for the original one in the solution path. We found that even the most advanced LMM such as GPT4o struggles at answering all related questions even if it answers the original question correctly, which is easy for human experts. A detailed case analysis is in Figure 6.

The phenomenon raises another concern for multiple-choice question (MCQ)-based multimodal evaluation benchmarks: correctly answering a question does not necessarily indicate that the model genuinely knows how to derive the final answer. A direct solution to accurately diagnose whether the model truly has the capability to solve the question is to let humans evaluate the model's reasoning process, as done in previous works like MathVista (Lu et al., 2024). However, human evaluation is labor-intensive and not reproducible, which complicates

the broader application of the method.

Several related works use advanced LLMs like GPT-4 to replace humans in evaluating the reasoning process of LMMs (Chen et al., 2024b; Zhang et al., 2024b). These methods show a strong correlation with human judgments in corresponding tests, but they result in unstable evaluation due to updates of proprietary models and inevitable API costs. Research in using LLMs as evaluators (Wang et al., 2023; Chen et al., 2024b) also finds systematic bias and a significant gap between open-source and proprietary models in terms of evaluation agreement with human experts. These drawbacks highlight the need for an economical, easy-to-use, and calibrated method for multimodal models.

3 MMEVALPRO: Calibrating Multimodal Evaluation

In this section, we delve into the detailed process of constructing the MMEVALPRO benchmark dataset and elucidate our methodologies for ensuring high-quality evaluation standards. Through these efforts, MMEVALPRO sets a new paradigm in the assessment of multimodal models, aiming to foster both accuracy and efficiency in multimodal evaluation.

3.1 Data Source

To enhance the diversity of our benchmark data, MMEVALPRO integrates content from three prominent multimodal benchmarks: MMMU, ScienceQA, and MathVista. These benchmarks span educational levels from junior high to undergraduate and cover various subjects. The details of the source datasets are in Appendix A.1. Considering annotator expertise and budget, we selected 328 questions from MMMU(dev), 1,200 from ScienceQA-Image, and all 540 from MathVista, totaling 2,138 distinct multimodal MCQs.

3.2 Annotation Pipeline

As illustrated in Figure 4, we design the following pipeline for MMEVALPRO to generate question triplets. The triplet consists of an original question, a perception question, and a knowledge question.

1) Data Preparation: Annotators begin by thoroughly reviewing the original question to ensure a deep understanding of concepts and solutions.

2) View and Analyse: Annotators are tasked with extracting crucial visual information and the logical framework implicit in the original problem, paving the way for the creation of nuanced perception and knowledge questions.

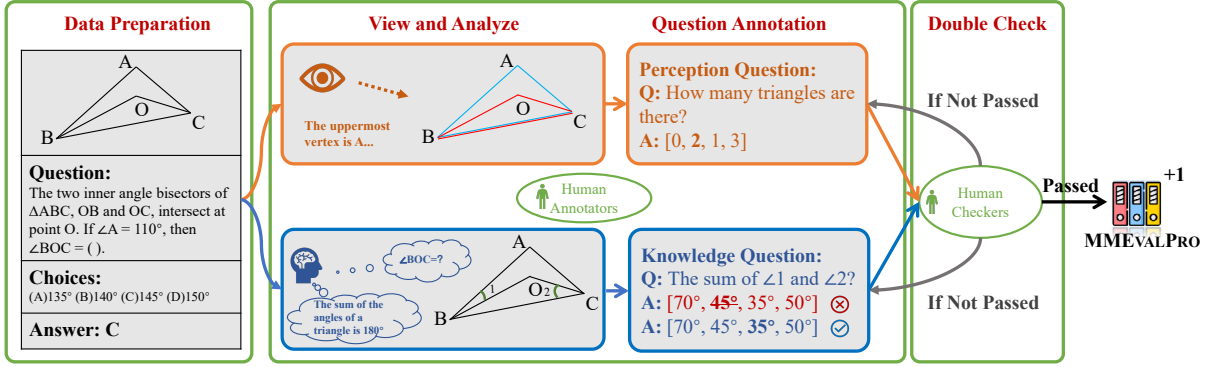


Figure 4: Annotation pipeline for MMEVALPRO.

3) Question Annotation: Building on the insights gathered, annotators then proceed to enrich the original question by formulating corresponding perception and knowledge questions, thereby expanding the scope of the evaluation.

4) Double Check: To maintain the integrity of the MMEVALPRO dataset, each annotated question triplet undergoes a rigorous verification process. Two independent checkers, who are not part of the annotation team, review each triplet for any errors or logical inconsistencies. Any issues identified prompt a re-annotation of the affected questions.

We provide the annotator guide in the supplement material. The final distribution of MMEVALPRO is shown in Figure 2. We list the key statistics of the benchmark in Table 4 from Appendix 3.1, the annotation guidelines in Appendix C and D.

3.3 Evaluation Metrics

We propose **Genuine Accuracy (GA)** as the primary metric for MMEVALPRO. GA equals 1 only if the model correctly answers the original question and the corresponding perception and knowledge prerequisite questions simultaneously. The second metric is the **Average Accuracy (AA)**, which computes the average accuracy of all questions.

MMEVALPRO can also be viewed as a multi-view evaluation process, where we naturally derive the **Perception Accuracy (PA)** score and the **Knowledge Accuracy (KA)** score by computing the average accuracy for the perception and knowledge anchor questions, respectively.

The **Consistency Gap (CG)** is measured by subtracting the Genuine Accuracy from the accuracy of the original question. This metric reflects the proportion of instances where the model correctly answers the original question but fails on more in-depth perception and knowledge questions, leading to inconsistency in its answers. To gain deeper

insights into the answer consistency and generalization capabilities of the tested models, we define **Perception Consistency (PC)** and **Knowledge Consistency (KC)** as the conditional probabilities $P(\text{Perception} = 1 \mid \text{Origin} = 1)$ and $P(\text{Knowledge} = 1 \mid \text{Origin} = 1)$, where $(= 1)$ indicates correctly answering the corresponding question in a question triplet. PC and KC together reveal the answer consistency of the tested model.

3.4 Quality Evaluation

To ensure the precision and consistency of our annotations, all question triplets are annotated by graduate students. For specialized college-level subjects in the MMMU sub-set, we hire students in the corresponding major to ensure annotation accuracy. All annotators are required to thoroughly familiarize themselves with the guide before commencing their annotation tasks. During the annotation process, we assign a minimum of three annotators for one question triplet. These foundations ensure the quality of the new annotated 4, 276 questions and there are only 47 questions that go through more than two double-check circles. Consensus was reached on all examples in the end.

4 Experiments

In this section, we evaluate a wide range of models on our MMEVALPRO benchmark. We first introduce the evaluation setup, and then present the quantitative results for both open-source and closed-source models. Finally we investigate the answer consistency gap among LLMs, LMMs and human experts with fine-grained analysis.

4.1 Setup

We test multiple LLMs and LMMs on the original benchmarks and MMEVALPRO. To streamline the evaluation process, all questions are converted

into a multiple-choice format, prompting the models to directly provide the answers. This approach simplifies answer matching process and eliminates the need for external models like ChatGPT, which is commonly used in previous studies such as (Lu et al., 2024; Chen et al., 2024b; Zhang et al., 2024b). In line with prior research, we incorporate an in-context demonstration for LLMs to standardize the output format. This is important as we have observed that different LLMs tend to produce varying response formats when not provided with an example. The detailed prompt and demonstration template are listed in Appendix B. For the original benchmarks, we report the average accuracy on the MCQ questions the same as we sampled for creating MMEVALPRO for fair comparison. For MMEVALPRO, we report both the Genuine Accuracy and Average Accuracy.

4.2 Evaluated Models

In our evaluation, we evaluate a variety of both LLMs and LMMs. For LLMs, we implement a 1-Shot setting, where a single demonstration is utilized to guide the output format. Among the LLMs assessed are some of the most advanced open-source models, including four versions of Qwen-1.5-Chat (Bai et al., 2023a) with sizes spanning from 7B to 72B, as well as the LLaMA3-Instruct (Touvron et al., 2023) series (8B and 70B). We also tested API-only models, such as GPT-3.5-Turbo (OpenAI, 2022) and GPT4-o (OpenAI, 2024), which rely solely on language. On the other hand, LMMs are evaluated in a zero-shot manner due to their ability to follow instructions to produce valid answer choices. In the open-source category, we tested the LLaVA-1.5 (Liu et al., 2023a) and LLaVA-Next (Liu et al., 2024) series, which include models of 7B, 13B, and 34B in size. We also evaluated two of the most cutting-edge LMMs, MiniCPM-V2.5-LLaMA3-8B (Hu et al., 2023) and InternVL-1.5-Chat-20B (Chen et al., 2023b), both of which have recently been made available to the public. For proprietary models, we tested GPT-4V-1106-preview (OpenAI, 2023), the latest GPT-4o (OpenAI, 2024), and Qwen-VL-Max (Bai et al., 2023b). We provide the prompts for LLMs and the hyperparameters used for LMMs in Appendix B.

4.3 Experiment Results

As shown in Table 2, we compare the performance of different LLMs and LMMs on MMEVALPRO and the original benchmarks. We evaluated the

performance of graduate students on the benchmarks as a strong baseline. The human evaluation guideline is shown in Appendix E. We also include random guess as a weak baseline performance.

We first evaluate LLMs and LMMs separately. For LLMs, all models perform poorly under the Genuine Accuracy metric in MMEVALPRO. For example, the advanced GPT-4o achieves 67.62% accuracy on the original ScienceQA-Image benchmark but drops to 10.30% on its calibrated version in MMEVALPRO, a decrease of 57.32 percentage points. Similar declines are observed in MMMU (down 37.38%) and MathVista (down 40%). These low scores more accurately reflect LLMs’ general multimodal capabilities due to their lack of visual perception abilities. Open-source LLMs also show drastic declines in Genuine Accuracy. The Qwen-1.5 series suggests that larger LLMs perform better on both benchmarks, supporting the idea that stronger LLMs are better at guessing. However, this trend does not hold for the LLaMA-3 series, indicating a need for our further investigation in the future. Generally, most LLMs score below 10% in Genuine Accuracy, highlighting the difficulty of the proposed benchmark for LLMs.

On the LMMs side, we also witness a large performance gap between the original benchmarks and MMEVALPRO. For proprietary models, GPT-4o and Qwen-VL-Max perform the best. If we compare the performance gap of best LLM and best LMM on the original benchmark and MMEVALPRO, we could find that the performance difference in scales is more clear on MMEVALPRO. For example, in original MMMU benchmark, GPT-4o with vision (54.85%) is only 1.1 times its non-vision LLM version (46.09%). While in MMEVALPRO, the LMM (25.08%) version’s performance is 2.9 times the LLM (8.71%) version of GPT-4o. A similar result also goes for other tasks. The enlarged performance discrepancy is more intuitive given the fundamental functionality difference between LLMs and LMMs. MMEVALPRO could better reflect the true abilities of examinees.

4.4 Fine-grained Analysis

To better explain models’ performance on MMEVALPRO, we conducted a fine-grained analysis on the experiments’ result according to the metrics proposed in section 3.3. We select the best performing open-source and proprietary LLMs and LMMs to analyze. The results are shown in Table 3. A detailed case analysis is shown in Figure 6.

Table 2: Main experiments result. For average, we report the macro average scores (the mean of different domains). Highest score is marked **bold** and the second highest is underlined. Best LMM still lags behind human with a substantial gap (-31.73% average **GA**) on MMEVALPRO.

Model	Open?	Original Source				MMEVALPRO			
		MMM	ScienceQA	MathVista	Average	MMM	ScienceQA	MathVista	Average
		(Average Accuracy)				(Genuine Accuracy / Average Accuracy)			
<i>Large Language Models (1-Shot)</i>									
Qwen-1.5-7B-Chat	Yes	34.48%	58.32%	39.07%	43.96%	1.95% / 30.30%	6.21% / 41.08%	2.78% / 29.19%	3.65% / 33.52%
Qwen-1.5-14B-Chat	Yes	32.68%	61.75%	43.14%	45.86%	5.48% / 35.05%	8.05% / 43.74%	4.63% / 35.74%	6.05% / 38.18%
Qwen-1.5-32B-Chat	Yes	39.21%	64.66%	42.96%	48.94%	5.81% / 39.56%	7.40% / 45.31%	5.56% / 37.53%	6.26% / 40.80%
Qwen-1.5-72B-Chat	Yes	39.33%	66.18%	43.52%	49.68%	5.80% / 38.06%	10.30% / 41.09%	5.19% / 36.98%	7.10% / 38.71%
LLaMA3-8B-Instruct	Yes	38.36%	68.57%	43.51%	50.15%	4.19% / 34.09%	6.22% / 45.24%	4.63% / 34.32%	5.01% / 37.88%
LLaMA3-70B-Instruct	Yes	35.56%	63.47%	37.59%	45.54%	4.84% / 34.94%	8.05% / 44.06%	2.59% / 31.85%	5.16% / 36.95%
GPT-3.5-Turbo	No	34.21%	61.85%	40.18%	45.41%	6.13% / 38.28%	9.98% / 45.87%	4.44% / 33.58%	6.85% / 39.24%
GPT-4o (non-vision)	No	46.09%	67.62%	45.19%	52.97%	8.71% / 39.90%	10.30% / 48.18%	5.19% / 37.78%	8.07% / 41.95%
<i>Large Multimodal Models (Zero-Shot)</i>									
LLaVA-1.5-Vicuna-7B	Yes	40.86%	62.61%	36.11%	46.53%	11.15% / 43.06%	25.64% / 58.33%	4.81% / 37.10%	13.87% / 46.16%
LLaVA-1.5-Vicuna-13B	Yes	39.92%	66.76%	37.59%	48.09%	10.82% / 43.28%	29.94% / 63.20%	5.19% / 36.98%	15.32% / 47.82%
LLaVA-Next-Vicuna-13B	Yes	39.75%	67.57%	41.48%	49.60%	12.13% / 47.86%	31.65% / 64.88%	9.07% / 39.26%	17.62% / 50.67%
LLaVA-Next-Hermes-Yi-34B	Yes	47.89%	74.77%	53.70%	58.79%	23.60% / 59.34%	43.67% / 73.35%	18.15% / 54.75%	28.47% / 62.48%
MiniCPM-V2.5-LLaMA3-8B	Yes	44.86%	77.68%	51.32%	57.95%	18.36% / 54.86%	43.56% / 73.96%	16.30% / 52.28%	26.07% / 60.37%
InternVL-1.5-Chat-20B	Yes	46.12%	87.27%	69.44%	67.61%	19.34% / 55.63%	45.49% / 76.25%	20.74% / 57.78%	28.52% / 63.22%
GPT-4V-1106-preview	No	44.32%	77.54%	51.85%	57.90%	22.30% / 57.95%	39.81% / 71.34%	13.04% / 48.48%	25.05% / 59.26%
GPT-4o	No	54.85%	<u>84.13%</u>	55.16%	<u>64.71%</u>	25.08% / 60.29%	49.68% / 76.86%	15.43% / 52.63%	<u>30.06%</u> / <u>63.26%</u>
Qwen-VL-Max	No	<u>48.75%</u>	76.22%	<u>59.07%</u>	61.35%	<u>24.38%</u> / <u>59.45%</u>	<u>47.61%</u> / 75.02%	21.48% / 59.19%	31.16% / 64.55%
Random Guess	-	26.04%	35.53%	32.41%	31.33%	1.94% / 28.60%	2.36% / 29.26%	3.32% / 29.14%	2.54% / 29.01%
Human (Graduate Student)	-	49.21%	76.92%	92.08%	72.74%	38.10% / 66.67%	64.42% / 81.09%	86.14% / 93.07%	62.89% / 80.28%

Table 3: Fine-grained scores on MMEVALPRO. **CG**: Consistency Gap, **PA**: Perception Accuracy, **KA**: Knowledge Accuracy, **PC**: Perception Consistency, **KC**: Knowledge Consistency

Model	Type	Open?	MMEVALPRO-MMMU					MMEVALPRO-ScienceQA					MMEVALPRO-MathVista				
			CG↓	PA↑	KA↑	PC↑	KC↑	CG↓	PA↑	KA↑	PC↑	KC↑	CG↓	PA↑	KA↑	PC↑	KC↑
Qwen-1.5-72b-Chat	LLM	Yes	33.53%	32.58%	41.29%	29.66%	41.52%	55.88%	20.17%	30.40%	30.32%	46.60%	38.33%	27.04%	44.43%	23.47%	45.07%
GPT-4o	LLM	No	37.38%	38.75%	43.24%	41.50%	41.51%	57.32%	46.87%	47.62%	46.43%	51.56%	40.00%	37.22%	38.15%	34.15%	41.46%
InternVL-1.5-Chat-20B	LMM	Yes	26.78%	63.87%	54.19%	63.57%	56.30%	41.78%	70.81%	71.46%	70.70%	70.95%	48.7%	51.67%	52.77%	52.69%	55.13%
GPT-4o	LMM	No	29.77%	63.07%	55.22%	63.33%	56.54%	34.45%	77.47%	75.21%	77.20%	75.39%	39.73%	57.32%	54.25%	53.63%	49.72%
Human	-	-	11.11%	80.65%	69.35%	93.56%	80.66%	12.50%	88.12%	83.17%	89.99%	88.77%	5.94%	94.06%	93.07%	95.70%	95.71%

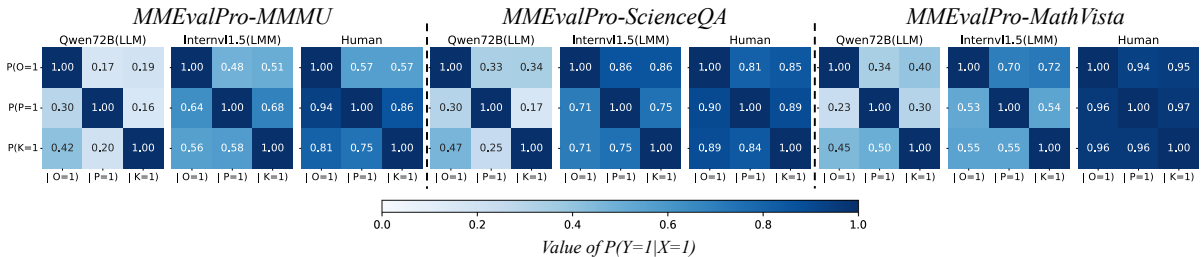


Figure 5: Heatmaps of conditional accuracy of MMEVALPRO.

More comparisons of different models are listed in Figure 16 from the Appendix.

Why MMEVALPRO is challenging and more trustworthy? When comparing human performance with that of LLMs and LMMs, we observe a more significant gap in MMEVALPRO evaluations than in original benchmarks. This indicates that MMEVALPRO is inherently a more challenging task. The primary difficulty stems from the issue of answer consistency, which demonstrates whether the model genuinely understands how to leverage perceptual abilities and knowledge to solve a given problem. To illustrate this, we compare the Consistency Gap (CG) scores among humans, the best-

performing LLMs and LMMs. The results suggest that LLMs generally exhibit a larger Consistency Gap than LMMs, while human experts display a considerably smaller CG. This trend is consistent across both open-source and proprietary models. A large Consistency Gap indicates that a model’s robustness and generalization abilities are limited: it may be able to answer the original question but fails to respond accurately to related prerequisite questions based on the same image.

This weakness is difficult to capture for the original benchmarks due to the single MCQ format. In fact, if we only consider the Average Accuracy of MMEVALPRO, as shown in Table 2, we ob-

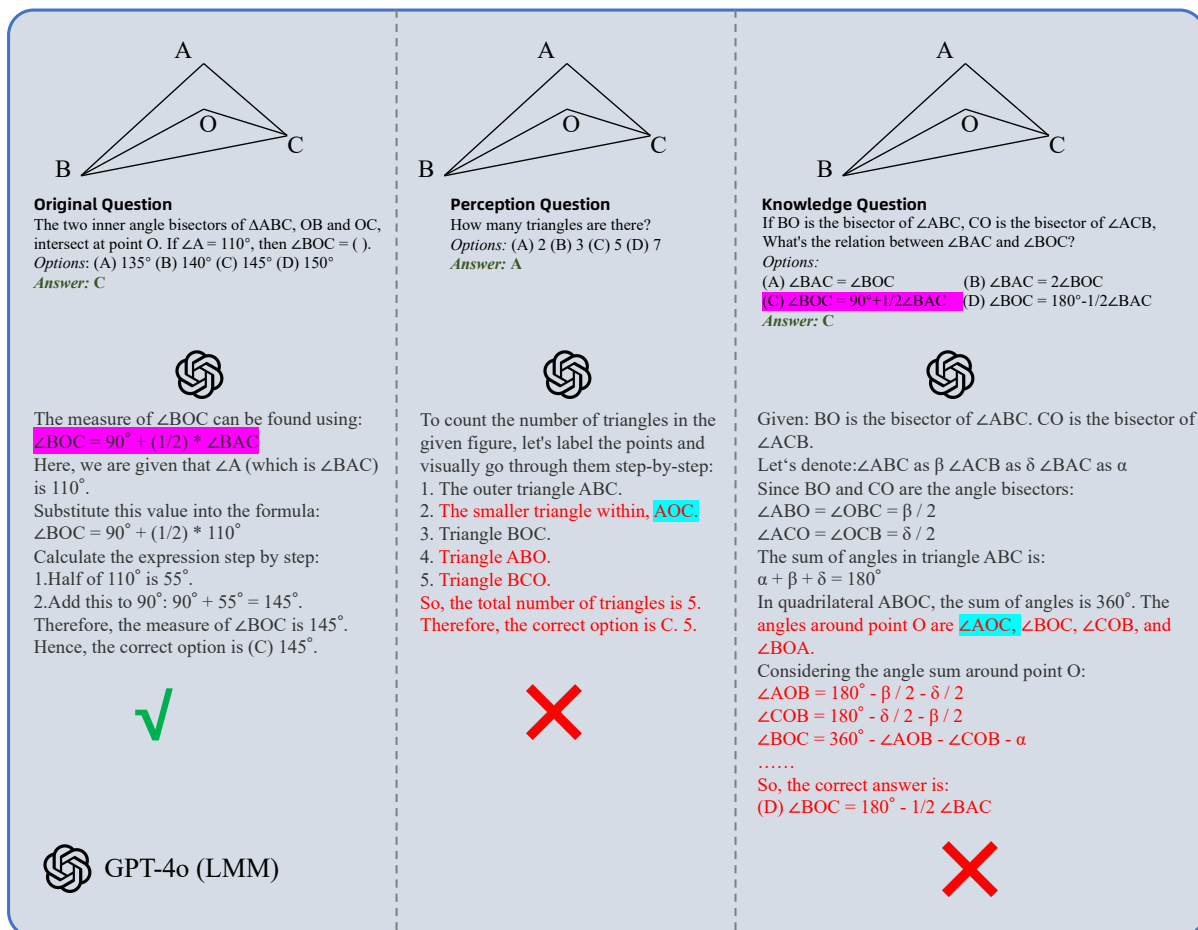


Figure 6: Case study on the answer inconsistency problem of LMMs. We could tell from the figure that GPT-4o could answer the original question however failed on the perception and knowledge question. The wrong reasoning process is marked red. The multiply mentioned items are marked with the same background color. In the reasoning process of original question, GPT-4o directly points out that $\angle BOC = 90^\circ + (1/2) \angle BAC$. However in the knowledge question while the " $\angle BOC = 90^\circ + (1/2) \angle BAC$ " is just one of the options, the model failed to figure it out, while human could easily achieve it. It shows the answer inconsistency problem of the LMM, which can not be pinpointed by single MCQ evaluation. If we further look at the reasoning process of the perception question, we could find that the model could not even recognize the correct number of triangles in the figure. For example it recognized AOC as a triangle, which is actually not. This also causes its problematic reasoning for the knowledge question.

serve that the performance difference compared to the original benchmarks is much smaller than the difference in Genuine Accuracy (GA). This suggests that LLMs can often guess correct answers for questions even without image input. MMEVAL-PRO addresses this issue effectively through the GA metric, making it a more reliable evaluation method compared to previous benchmarks.

Why the Consistency Gap is large? The gap between Genuine Accuracy and Average Accuracy on the original benchmarks reveals the answer inconsistency problem. We are further interested in what causes the problem. There are two possible reasons for a large Consistency Gap, that the model correctly answer the original question however fails

on the perception or knowledge one. We compare the **Perception Consistency (PC)** and **Knowledge Consistency (KC)** of the evaluated models and humans. We find that there is a clear performance border between humans and LMMs, LMMs and LLMs. Humans could reach at least 90%PC and 80%KC in various sub-tasks, showing strong answer consistency. While the numbers for LMMs are 50%PC and 55%KC, for LLMs are 23%PC and 41%KC according to Table 3.

PC and KC intuitively reflect the model's likelihood of correctly answering perception and knowledge questions if it has already solved the original question. Low PC and KC scores lead to a significant consistency gap. Beyond PC and KC, we visualize all conditional probabilities for the best

open-source LLM and LMM in Figure 5. Humans generally exhibit high probabilities of correctly answering one question given a correct answer to another, indicating consistent thinking. In contrast, LLMs show the lowest probabilities compared to LMMs and humans. This is expected, as LLMs lack consistent multimodal problem-solving paths due to their absence of visual perception, thus supporting the credibility of the benchmarks.

Examining the Perception Accuracy (PA) and Knowledge Accuracy (KA) in Table 3, we find that LMMs demonstrate a greater advantage in PA compared to KA when contrasted with LLMs. This is because PA depends directly on visual capabilities, which LLMs lack. The above conclusion explains why tested models have larger CG compared to humans, which is a potential and promising direction for future LMMs to improve on.

5 Related Work

There have been several benchmarks built for evaluating LMMs (Feng et al., 2024; Yang et al., 2024), such as MMBench, MME, Seed-Bench (Liu et al., 2023c; Fu et al., 2023; Li et al., 2023a) that assess LMMs performance from multiple fine-grained dimensions. LVLM-eHub, M3IT (Xu et al., 2023; Li et al., 2023b) focus on the general instruction following ability. MMMU, MathVista, ScienceQA (Yue et al., 2023; Lu et al., 2024, 2022) require perception from the vision part and knowledge in the language part.

Nonetheless, critiques have been raised regarding the limitations of these existing benchmarks in effectively evaluating LMMs. PCA-Bench, MathVerse (Chen et al., 2024b; Zhang et al., 2024b) adopt strong LLMs such as GPT4 and GPT4-Vision to score the reasoning process of LMMs in embodied-AI and math diagram questions, in order to pinpoint cases where the LMM gets the correct answer by a fluke. Yet, using a proprietary model to conduct evaluation hinders the broader usage of the method, moreover, the evaluation result has bias itself due to the proxy model and would change over time. MMStar (Chen et al., 2024d) filters out the questions that do not rely on visual information in existing multimodal benchmarks. However, it does not address the issue inherent in MCQ, where models can potentially get the correct answer without truly understanding the content. Compared with those benchmarks, MMEVALPRO is more economical, easy-to-use, and calibrated for

evaluating multimodal models.

6 Conclusion

We propose MMEVALPRO, a multimodal benchmark designed to address issues identified in previous evaluations and built upon MMMU, ScienceQA-Image, and MathVista. MMEVALPRO introduces twin perception and knowledge anchors to the original framework and defines Genuine Accuracy as its primary metric, thereby reducing the likelihood of LLMs manipulating the questions. Our extensive experiments and analyses on a wide array of models and human experts demonstrate that MMEVALPRO more accurately reflects the true capabilities of the tested LMMs and presents a more challenging task. Notably, even the most advanced models, such as GPT-4o and Qwen-VL-Max, trail behind human performance by a substantial gap of more than 30% in Genuine Accuracy. Our analysis into the reasons behind the consistency gap problem elucidates the disparity and provides valuable insights for future research.

Limitations

In order to ensure the high quality and accuracy of MMEVALPRO, we employed manual annotation with human experts to construct the dataset. A certain level of human effort and expertise are necessary. To some extent, this requirement may limit the expansion efficiency of MMEVALPRO.

Acknowledgments

This paper is partially supported by the National Key Research and Development Program of China with Grant No.2023YFC3341203 as well as the National Natural Science Foundation of China with Grant Numbers 61876004 and 62306014.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. [Qwen technical report](#). *Preprint*, arXiv:2309.16609.

- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *ArXiv preprint*, abs/2308.12966.
- Liang Chen, Zekun Wang, Shuhuai Ren, Lei Li, Haozhe Zhao, Yunshui Li, Zefan Cai, Hongcheng Guo, Lei Zhang, Yizhe Xiong, et al. 2024a. Next token prediction towards multimodal intelligence: A comprehensive survey. *arXiv preprint arXiv:2412.18619*.
- Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Tianyu Liu, and Baobao Chang. 2023a. Towards end-to-end embodied decision making via multi-modal large language model: Explorations with gpt4-vision and beyond. *ArXiv*.
- Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Xiangdi Meng, Tianyu Liu, and Baobao Chang. 2024b. *Pca-bench: Evaluating multimodal large language models in perception-cognition-action chain*. *Preprint*, arXiv:2402.15527.
- Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024c. *An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models*. *Preprint*, arXiv:2403.06764.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024d. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.
- Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.
- Colin B. Clement, Matthew Bierbaum, Kevin P. O’Keefe, and Alexander A. Alemi. 2019. *On the use of arxiv as a dataset*. *Preprint*, arXiv:1905.00075.
- Bin Feng, Zequn Liu, Nanlan Huang, Zhiping Xiao, Haomiao Zhang, Srubhi Mirzoyan, Hanwen Xu, Jiaran Hao, Yinghui Xu, Ming Zhang, et al. 2024. A bioactivity foundation model using pairwise meta-learning. *Nature Machine Intelligence*, 6(8):962–974.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. *Mme: A comprehensive evaluation benchmark for multimodal large language models*. *Preprint*, arXiv:2306.13394.
- Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038*.
- Aniruddha Kembhavi, Michael Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. *A diagram is worth a dozen images*. *ArXiv*, abs/1603.07396.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. *Seed-bench: Benchmarking multimodal llms with generative comprehension*. *Preprint*, arXiv:2307.16125.
- Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023b. *M³IT: A large-scale dataset towards multi-modal multilingual instruction tuning*. *ArXiv preprint*, abs/2306.04387.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. *Llava-next: Improved reasoning, ocr, and world knowledge*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *ArXiv preprint*, abs/2304.08485.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023c. *Mmbench: Is your multi-modal model an all-around player?* *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. 2024. *Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts*. *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. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.
- OpenAI. 2022. Introducing chatgpt. *Web*.
- OpenAI. 2023. Gpt-4v(ision) system card. *Web*.
- OpenAI. 2024. *hello-gpt-4o*.
- Jianhao Shen, Ye Yuan, Srubhi Mirzoyan, Ming Zhang, and Chenguang Wang. 2024. *Measuring vision-language stem skills of neural models*. *Preprint*, arXiv:2402.17205.

Gemini Team, Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry, Lepikhin, Timothy Lillcrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, Luke Vilnis, Oscar Chang, Nobuyuki Morioka, George Tucker, Ce Zheng, Oliver Woodman, Nithya Ataluri, Tomas Kocisky, Evgenii Eltyshev, Xi Chen, Timothy Chung, Vittorio Selo, Siddhartha Brahma, Petko Georgiev, Ambrose Slone, Zhenkai Zhu, James Lottes, Siyuan Qiao, Ben Caine, Sebastian Riedel, Alex Tomala, Martin Chadwick, Juliette Love, Peter Choy, Sid Mittal, Neil Houlsby, Yunhao Tang, Matthew Lamm, Libin Bai, Qiao Zhang, Luheng He, Yong Cheng, Peter Humphreys, Yujia Li, Sergey Brin, Albin Cassirer, Yingjie Miao, Lukas Zilka, Taylor Tobin, Kelvin Xu, Lev Proleev, Daniel Sohn, Alberto Magni, Lisa Anne Hendricks, Isabel Gao, Santiago Ontanon, Oskar Bunyan, Nathan Byrd, Abhanshu Sharma, Biao Zhang, Mario Pinto, Rishika Sinha, Harsh Mehta, Dawei Jia, Sergi Caelles, Albert Webson, Alex Morris, Becca Roelofs, Yifan Ding, Robin Strudel, Xuehan Xiong, Marvin Ritter, Mostafa Dehghani, Rahma Chaabouni, Abhijit Karmarkar, Guangda Lai, Fabian Mentzer, Bibo Xu, YaGuang Li, Yujing Zhang, Tom Le Paine, Alex Goldin, Behnam Neyshabur, Kate Baumli, Anselm Levskaya, Michael Laskin, Wenhao Jia, Jack W. Rae, Kefan Xiao, Antoine He, Skye Giordano, Lakshman Yagati, Jean-Baptiste Lespiau, Paul Natsev, Sanjay Ganapathy, Fangyu Liu, Danilo Martins, Nanxin Chen, Yunhan Xu, Megan Barnes, Rhys May, Arpi Vezer, Junhyuk Oh, Ken Franko, Sophie Bridgers, Ruizhe Zhao, Boxi Wu, Basil Mustafa, Sean Sechrist, Emilio Parisotto, Thanumalayan Sankaranarayanan Pillai, Chris Larkin, Chenjie Gu, Christina Sorokin, Maxim Krikun, Alexey Guseynov, Jessica Landon, Romina Datta, Alexander Pritzel, Phoebe Thacker, Fan Yang, Kevin Hui, Anja Hauth, Chih-Kuan Yeh, David Barker, Justin Mao-Jones, Sophia Austin, Hannah Sheahan, Parker Schuh, James Svensson, Rohan Jain, Vinay Ramasesh, Anton Briukhov, DaWoon Chung, Tamara von Glehn, Christina Butterfield, Priya Jhakra, Matthew Wiethoff, Justin Frye, Jordan Grimstad, Beer Changpinyo, Charline Le Lan, Anna Bortsova, Yonghui Wu, Paul Voigtlaender, Tara Sainath, Shane Gu, Charlotte Smith, Will Hawkins, Kris Cao, James Besley, Srivatsan Srinivasan, Mark Omernick, Colin Gaffney, Gabriela Surita, Ryan Burnell, Bogdan Damoc, Junwhan Ahn, Andrew Brock, Mantas Pajarskas, Anastasia Petrushkina, Seb Noury, Lorenzo Blanco, Kevin

Swersky, Arun Ahuja, Thi Avrahami, Vedant Misra, Raoul de Liedekerke, Mariko Iinuma, Alex Polozov, Sarah York, George van den Driessche, Paul Michel, Justin Chiu, Rory Blevins, Zach Gleicher, Adrià Recasens, Alban Rustemi, Elena Gribovskaya, Aurko Roy, Wiktor Gworek, Sébastien M. R. Arnold, Lisa Lee, James Lee-Thorp, Marcello Maggioni, Enrique Piqueras, Kartikeya Badola, Sharad Vikram, Lucas Gonzalez, Anirudh Baddepudi, Evan Senter, Jacob Devlin, James Qin, Michael Azzam, Maja Trebacz, Martin Polacek, Kashyap Krishnakumar, Shuo yiin Chang, Matthew Tung, Ivo Penchev, Rishabh Joshi, Kate Olszewska, Carrie Muir, Mateo Wirth, Ale Jakse Hartman, Josh Newlan, Sheleem Kashem, Vijay Bolina, Elahe Dabir, Joost van Amersfoort, Zafarali Ahmed, James Cobon-Kerr, Aishwarya Kamath, Arnar Mar Hrafnkelsson, Le Hou, Ian Mackinnon, Alexandre Frechette, Eric Noland, Xiance Si, Emanuel Taropa, Dong Li, Phil Crone, Anmol Gulati, Sébastien Cevey, Jonas Adler, Ada Ma, David Silver, Simon Tokumine, Richard Powell, Stephan Lee, Kiran Vodrahalli, Samer Hassan, Diana Mincu, Antoine Yang, Nir Levine, Jenny Brennan, Mingqiu Wang, Sarah Hodkinson, Jeffrey Zhao, Josh Lipschultz, Aedan Pope, Michael B. Chang, Cheng Li, Laurent El Shafey, Michela Paganini, Sholto Douglas, Bernd Bohnet, Fabio Pardo, Seth Odoom, Mihaela Rosca, Cicero Nogueira dos Santos, Kedar Soparkar, Arthur Guez, Tom Hudson, Steven Hansen, Chulayuth Asawaroengchai, Ravi Addanki, Tianhe Yu, Wojciech Stokowiec, Mina Khan, Justin Gilmer, Jaehoon Lee, Carrie Grimes Bostock, Keran Rong, Jonathan Caton, Pedram Pejman, Filip Pavetic, Geoff Brown, Vivek Sharma, Mario Lučić, Rajkumar Samuel, Josip Djolonga, Amol Mandhane, Lars Lowe Sjöstrand, Elena Buchatskaya, Elspeth White, Natalie Clay, Jiepu Jiang, Hyeontaek Lim, Ross Hemsley, Zeynep Cankara, Jane Labanowski, Nicola De Cao, David Steiner, Sayed Hadi Hashemi, Jacob Austin, Anita Gergely, Tim Blyth, Joe Stanton, Kaushik Shivakumar, Aditya Siddhant, Anders Andreassen, Carlos Araya, Nikhil Sethi, Rakesh Shivanna, Steven Hand, Ankur Bapna, Ali Khodaei, Antoine Miech, Garrett Tanzer, Andy Swing, Shantanu Thakoor, Lora Aroyo, Zhufeng Pan, Zachary Nado, Jakub Sygnowski, Stephanie Winkler, Dian Yu, Mohammad Saleh, Loren Maggiore, Yamini Bansal, Xavier Garcia, Mehran Kazemi, Piyush Patil, Ishita Dasgupta, Iain Barr, Minh Giang, Thais Kagohara, Ivo Danihelka, Amit Marathe, Vladimir Feinberg, Mohamed Elhawaty, Nimesh Ghelani, Dan Horgan, Helen Miller, Lexi Walker, Richard Tanburn, Mukarram Tariq, Disha Shrivastava, Fei Xia, Qingze Wang, Chung-Cheng Chiu, Zoe Ashwood, Khuslen Baatarsukh, Sina Samangooei, Raphaël Lopez Kaufman, Fred Alcober, Axel Stjerngren, Paul Komarek, Katerina Tsihlas, Anudhyan Boral, Ramona Comanescu, Jeremy Chen, Ruiho Liu, Chris Welty, Dawn Bloxwich, Charlie Chen, Yanhua Sun, Fangxiaoyu Feng, Matthew Mauer, Xerxes Dotiwalla, Vincent Hellendoorn, Michael Sharman, Ivy Zheng, Krishna Haridasan, Gabe Barth-Maron, Craig Swanson, Dominika Rogozińska, Alek Andreev, Paul Kishan Rubenstein, Ruoxin Sang, Dan Hurt, Gamaleldin Elsayed, Ren-

shen Wang, Dave Lacey, Anastasija Ilić, Yao Zhao, Adam Iwanicki, Alejandro Lince, Alexander Chen, Christina Lyu, Carl Lebsack, Jordan Griffith, Meenu Gaba, Paramjit Sandhu, Phil Chen, Anna Koop, Ravi Rajwar, Soheil Hassas Yeganeh, Solomon Chang, Rui Zhu, Soroush Radpour, Elnaz Davoodi, Ving Ian Lei, Yang Xu, Daniel Toyama, Constant Segal, Martin Wicke, Hanzhao Lin, Anna Bulanova, Adrià Puigdomènech Badia, Nemanja Rakićević, Pablo Sprechmann, Angelos Filos, Shaobo Hou, Víctor Campos, Nora Kassner, Devendra Sachan, Meire Fortunato, Chimezie Iwuanyanwu, Vitaly Nikolaev, Balaji Lakshminarayanan, Sadegh Jazayeri, Mani Varadarajan, Chetan Tekur, Doug Fritz, Misha Khalman, David Reitter, Kingshuk Dasgupta, Shourya Sarcar, Tina Ornduff, Javier Snaider, Fantine Huot, Johnson Jia, Rupert Kemp, Nejc Trdin, Anitha Vijayakumar, Lucy Kim, Christof Angermueller, Li Lao, Tianqi Liu, Haibin Zhang, David Engel, Somer Greene, Anaïs White, Jessica Austin, Lilly Taylor, Shereen Ashraf, Danyi Liu, Maria Georgaki, Irene Cai, Yana Kulizhskaya, Sonam Goenka, Brennan Saeta, Ying Xu, Christian Frank, Dario de Cesare, Brona Robenek, Harry Richardson, Mahmoud Alnahlawi, Christopher Yew, Priya Ponnappalli, Marco Tagliasacchi, Alex Korchemniy, Yelin Kim, Dinghua Li, Bill Rosgen, Kyle Levin, Jeremy Wiesner, Praseem Banzal, Praveen Srinivasan, Hongkun Yu, Çağlar Ünlü, David Reid, Zora Tung, Daniel Finchelstein, Ravin Kumar, Andre Elisseeff, Jin Huang, Ming Zhang, Ricardo Aguilar, Mai Giménez, Jiawei Xia, Olivier Dousse, Willi Gierke, Damion Yates, Komal Jalan, Lu Li, Eri Latorre-Chimoto, Duc Dung Nguyen, Ken Durden, Praveen Kallakuri, Yaxin Liu, Matthew Johnson, Tomy Tsai, Alice Talbert, Jasmine Liu, Alexander Neitz, Chen Elkind, Marco Selvi, Mimi Jasarevic, Livio Baldini Soares, Albert Cui, Pidong Wang, Alek Wenjiao Wang, Xinyu Ye, Krystal Kallarackal, Lucia Loher, Hoi Lam, Josef Broder, Dan Holtmann-Rice, Nina Martin, Bramandia Ramadhana, Mrinal Shukla, Sujoy Basu, Abhi Mohan, Nick Fernando, Noah Fiedel, Kim Paterson, Hui Li, Ankush Garg, Jane Park, DongHyun Choi, Diane Wu, Sankalp Singh, Zhishuai Zhang, Amir Globerson, Lily Yu, John Carpenter, Félix de Chaumont Quitry, Carey Radebaugh, Chu-Cheng Lin, Alex Tudor, Prakash Shroff, Drew Garmon, Dayou Du, Neera Vats, Han Lu, Shariq Iqbal, Alex Yakubovich, Nilesh Tripuraneni, James Manyika, Haroon Qureshi, Nan Hua, Christel Ngani, Maria Abi Raad, Hannah Forbes, Jeff Stanway, Mukund Sundararajan, Victor Ungureanu, Colton Bishop, Yunjie Li, Balaji Venkatraman, Bo Li, Chloe Thornton, Salvatore Scellato, Nishesh Gupta, Yicheng Wang, Ian Tenney, Xihui Wu, Ashish Shenoy, Gabriel Carvajal, Diana Gage Wright, Ben Bariach, Zhuyun Xiao, Peter Hawkins, Sid Dalmia, Clement Farabet, Pedro Valenzuela, Quan Yuan, Ananth Agarwal, Mia Chen, Wooyeol Kim, Brice Hulse, Nandita Dukkipati, Adam Paszke, Andrew Bolt, Kiam Choo, Jennifer Beattie, Jennifer Prendki, Harsha Vashisht, Rebeca Santamaria-Fernandez, Luis C. Cobo, Jarek Wilkiewicz, David Madras, Ali Elqursh, Grant Uy, Kevin Ramirez, Matt Harvey, Tyler Liechty, Heiga Zen, Jeff Seibert,

Clara Huiyi Hu, Andrey Khorlin, Maigo Le, Asaf Aharoni, Megan Li, Lily Wang, Sandeep Kumar, Norman Casagrande, Jay Hoover, Dalia El Badawy, David Soergel, Denis Vnukov, Matt Miecznikowski, Jiri Simsa, Praveen Kumar, Thibault Sellam, Daniel Vlasic, Samira Daruki, Nir Shabat, John Zhang, Guolong Su, Jiageng Zhang, Jeremiah Liu, Yi Sun, Evan Palmer, Alireza Ghaffarkhah, Xi Xiong, Victor Cotruta, Michael Fink, Lucas Dixon, Ashwin Sreevatsa, Adrian Goedeckemeyer, Alek Dimitriev, Mohsen Jafari, Remi Crocker, Nicholas FitzGerald, Aviral Kumar, Sanjay Ghemawat, Ivan Philips, Frederick Liu, Yannie Liang, Rachel Sterneck, Alena Repina, Marcus Wu, Laura Knight, Marin Georgiev, Hyo Lee, Harry Askham, Abhishek Chakladar, Annie Louis, Carl Crous, Hardie Cate, Dessie Petrova, Michael Quinn, Denese Owusu-Afriyie, Achintya Singhal, Nan Wei, Solomon Kim, Damien Vincent, Milad Nasr, Christopher A. Choquette-Choo, Reiko Tojo, Shawn Lu, Diego de Las Casas, Yucheng Cheng, Tolga Bolukbasi, Katherine Lee, Saaber Fatehi, Rajagopal Ananthanarayanan, Miteyan Patel, Charbel Kaed, Jing Li, Shreyas Rammohan Belle, Zhe Chen, Jaclyn Konzelmann, Siim Pöder, Roopal Garg, Vinod Koverkathu, Adam Brown, Chris Dyer, Rosanne Liu, Azade Nova, Jun Xu, Alanna Walton, Alicia Parrish, Mark Epstein, Sara McCarthy, Slav Petrov, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2024. [Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context](#). *Preprint*, arXiv:2403.05530.

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

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. [Large language models are not fair evaluators](#). *ArXiv*, abs/2305.17926.

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](#). *Preprint*, arXiv:2306.09265.

Junwei Yang, Hanwen Xu, Srubhi Mirzoyan, Tong Chen, Zixuan Liu, Zequn Liu, Wei Ju, Luchen Liu, Zhiping Xiao, Ming Zhang, et al. 2024. [Poisoning medical knowledge using large language models](#). *Nature Machine Intelligence*, 6(10):1156–1168.

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 Wenhao Chen. 2023. [Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi](#). *Preprint*, arXiv:2311.16502.

Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, Haoran Zhang, Xingwei Qu, Junjie Wang, Ruibin Yuan, Yizhi Li, Zekun Wang, Yudong Liu, Yu-Hsuan Tsai, Fengji Zhang, Chenghua Lin, Wenhao Huang, Wenhui Chen, and Jie Fu. 2024a. *Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark*. *Preprint*, arXiv:2401.11944.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024b. *Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?* *Preprint*, arXiv:2403.14624.

Appendix

A Details on MMEVALPRO

A.1 Data Source

MMMU (Yue et al., 2023): The MMMU is a benchmark designed to evaluate multimodal models on massive multi-discipline tasks. The benchmark sources its questions from college examinations, quizzes, and textbooks, encompassing 30 subjects and 183 subfields across six core disciplines, i.e. Art & Design, Business, Science, Health & Medicine, Humanities & Social Science and Tech & Engineering. This benchmark is meticulously designed to assess models’ capabilities in handling multi-disciplinary tasks, drawing upon college-level subject knowledge.

ScienceQA (Lu et al., 2022): The Science Question Answering (ScienceQA) is another pivotal resource, which consists of 21,208 multimodal multiple choice questions with diverse science topics. There are only 48.7% questions of ScienceQA that have an image context. It is renowned for its application in multimodal tasks and features a domain diversity spanning three primary science subjects, i.e., natural, language, and social. The dataset comprises multimodal science questions that are collated from elementary and high school science curricula, ensuring a breadth of scientific inquiry and comprehension. The subjects of the questions can be categorized by Biology, Physics, Chemistry, and others.

MathVista (Lu et al., 2024): The MathVista benchmark is developed to evaluate the reasoning ability of the multimodal models, which consists of 6,141 examples from 31 datasets (28 mathematics and IQTest, FunctionQA, PaperQA). The dataset offers exclusively mathematical and visual tasks. This source enriches our dataset with rigorous computational and analytical problems, providing a robust framework for evaluating quantitative reasoning in multimodal contexts.

Finally, we select the validation set of MMMU (722 questions), the questions with images in ScienceQA (2,097 questions), and the testmini set of MathVista (540 questions) to construct the MMEVALPRO. From the 722 questions in the validation set of MMMU, we chose the questions with topics suited for the annotator’s major and other questions in easy-level to annotate, which resulted in 339 final questions. And we annotated all questions selected from the testmini set of MathVista

and 1, 259 of ScienceQA-Image.

The distribution of MMEVALPRO is shown in the Figure 2. There are 58.89% questions originating from ScienceQA, 25.26% from MathVista, and 15.86% from MMMU. We further categorize the problems by task categories, subjects, and domains, computing their relative percentages to the total questions. In the annotated questions originated from ScienceQA, there are 16.68% of them subject to Geography, 12.61% subject to Physics, 12.02% with the Biology topic, and the remaining questions distributed in History, Chemistry, Economics, etc. In MathVista, the annotated questions contain 5 tasks, including GPS(geometry problem solving), VQA(visual question answering), FQA(figure question answering), TQA(textbook question answering), and MWP(math word problem), the last one only accounted for a very small proportion and is not annotated in the figure. According to the subjects of questions from MMMU, there are 4.91% questions in Medicine, 3.27% in Art, 3.65% in Business, 2.71% in Science, and 1.32% from other subjects with easy-level.

A.2 License

We check all the datasets' licenses and they all permit customization and redistribution for non-commercial use. MMMU is under Apache 2.0 License, ScienceQA is under MIT License and MathVista is covered by CC BY-SA 4.0.

MMEVALPRO can be used commercially as a test set, but using it as a training set is prohibited. By accessing or using this dataset, users acknowledge and agree to abide by these terms in conjunction with the CC BY-SA 4.0 license. We make sure there is no offensive content found during the whole annotation process.

A.3 MMEVALPRO Examples

We list examples of different splits of MMEVALPRO as shown in Figure 7, 8 and 9.

A.4 Statistics of MMEVALPRO

The key statistics of MMEVALPRO is shown in the table 4. Figure 10 display the distribution of answers and the number of options in the MMEVALPRO.

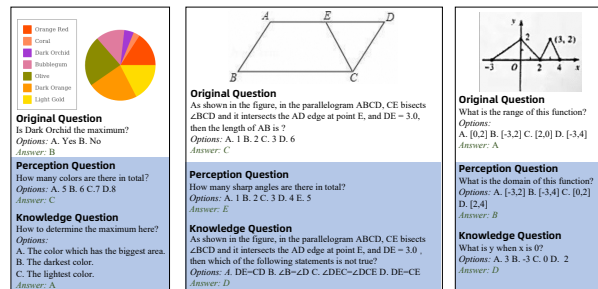


Figure 7: Examples of the MathVista subset of MMEVALPRO.

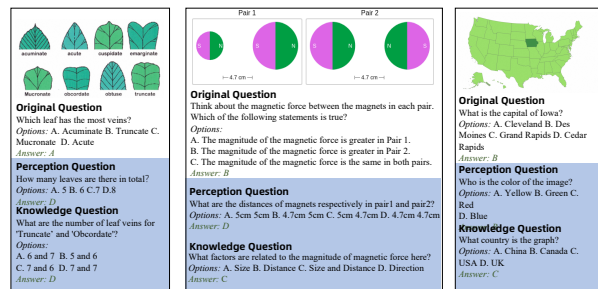


Figure 8: Examples of the ScienceQA subset of MMEVALPRO.

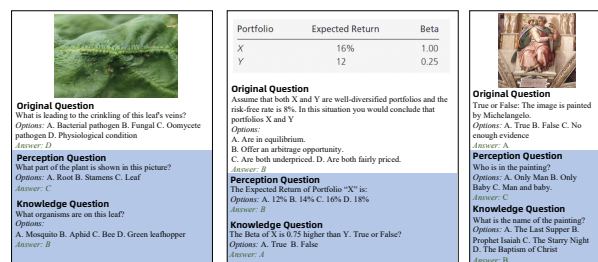


Figure 9: Examples of the MMMU subset of MMEVALPRO.

Statistic	Number
Source datasets	3
Number of question triplets	2,138
Number of unique questions	6,414
Triplets from MMMU	339
Triplets from ScienceQA	1259
Triplets from MathVista	540
Maximum question length	165
Maximum choice number	12
Average question length	9.60
Average choice number	3.94

Table 4: Key statistics of MMEVALPRO.

B Experiment Setup

B.1 Prompt Format for Different Models

For LLM "Given a question you need to choose the best answer from the given options. I will first

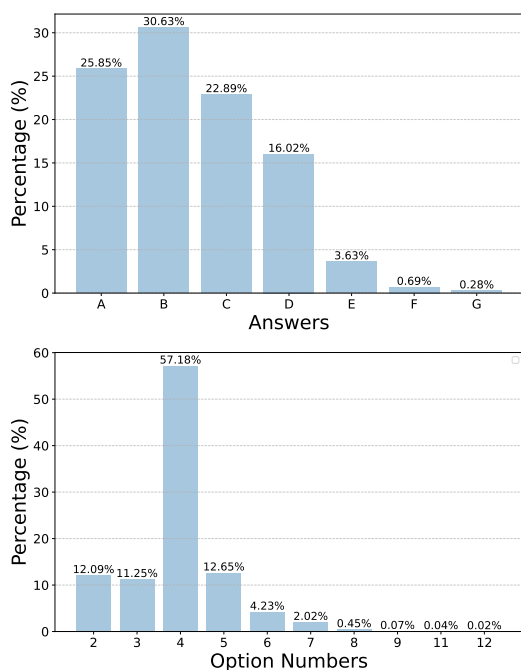


Figure 10: Distribution of Answer Choices in MMEVALPRO.

give you an example, you need to follow the output format of the answer. Example: Question: {example question} Options: {example options} Output: {example answer}. Just answer the following question with only the letter of the correct option, or you will get no credit. Question:{question} Options: {options}"

We use a fixed demonstration for all inferences to format the output. The example is: "**Question:** Baxter Company has a relevant range of production between 15,000 and 30,000 units. The following cost data represents the average variable costs per unit for 25,000 units of production. If 30,000 units are produced, what are the per unit manufacturing overhead costs incurred? **Options:** (A) \$6 (B) \$7 (C) \$8 (D) \$9 **Output:** A".

For LMM "Analyse the image and choose the best answer for the following question:{question} Options: {options} Just output the letter of the correct answer."

B.2 Model Hyper-parameters

For open-source models, we use the default inference script provided in corresponding papers and githubs. For proprietary models, we follow the official guide to call the API. In particular, we do not use sampling techniques during generation to ensure our results are reproducible. All experiments

are done on a local server with 4 NVIDIA-A100 GPUs.

C MMEVALPRO Triplet Annotation Guideline

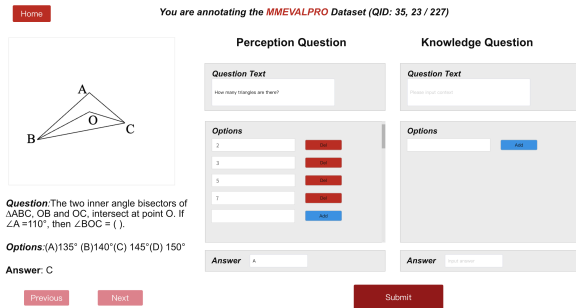


Figure 11: UI for annotations of MMEVALPRO.

Our research requires the construction of a more trustable multimodal evaluation dataset. The collection and arrangement of original questions have been completed. You, as annotators, need to perform data annotations based on the original questions, expanding the original question into a set of three questions (original, perception, knowledge). Every annotators should carefully read the following content before embarking on the annotation task.

- The data annotation task will be completed on the annotation UI interface we have built.
 - Each annotator will be allocated an account and password.
 - Each account corresponds to an independent subset of questions needing annotation.
- Each original question to be annotated includes an image, a question text, possible answers, and the correct answer. You need to **read the original question** and **understand the logic of solution** firstly.
- You need to provide two entirely new questions, a **perceptual question**, and a **knowledge question** based on step 2.
 - The perceptual question should be a question **related to the content of the image** corresponding to the original question.
 - The knowledge question should be **related to the logic of solving** the original question.
 - You should input **the text of the question, the variable options, and the correct answer**. Each question must be a **multiple-choice question**. Once confirmed to be correct, you can click to submit.

Figure 12: Annotation Guideline of MMEVALPRO.

We developed an annotation Web UI to enable expert annotators to construct question triplets of MMEVALPRO. The Web UI is shown in the figure 11. The annotators of MMEVALPRO were trained with the guideline shown in figure 12 before formal work.

D MMEVALPRO Triplet Checking Guideline

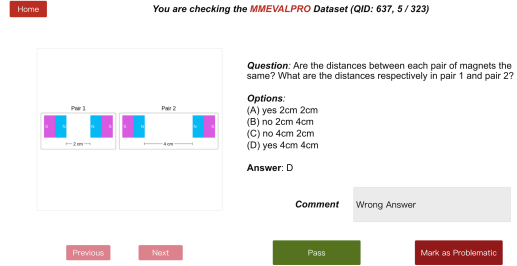


Figure 13: UI for checking of MMEVALPRO.

Our research requires the construction of a more trustable multimodal evaluation dataset. The data annotation process has been completed. The reviewers need to check the questions annotated by the annotators, and see if there are any issues with the corresponding question descriptions and the provided correct answers. Please read the following content before starting your review.

- The review process of the dataset will be carried out on a dedicated webpage. You will be assigned a corresponding account and password.
- During the review of the questions, please check all the content annotated by the annotators, including:
 - Question Text** (Ambiguous description, mismatch with the image, grammatical errors, etc.)
 - Question Options** (Duplication, ambiguity, and other problems.)
 - Answer** (Is the answer correct?)
- You need to mark the questions without problems as **Pass** and mark problematic questions as **Problematic**. Appropriate comments also need to be made on the problematic questions for the data annotator to correct (for example, "ambiguous", "the answer is incorrect", etc.)

Figure 14: Checking Guideline of MMEVALPRO.

In our study, we employed double checking to maintain the quality of MMEVALPRO. We also developed a web page shown in figure 13 for checking. The checkers were trained with checking instructions shown in figure 14.

E Human Evaluation Guide

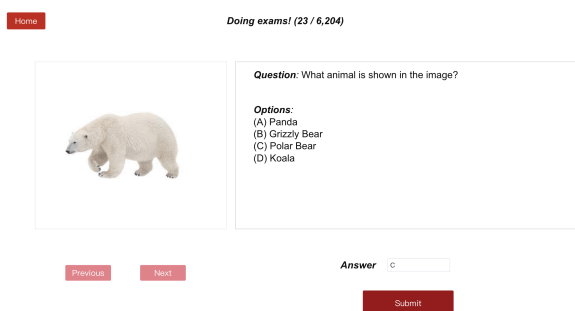


Figure 15: UI for human evaluation of MMEVALPRO.

To ensure the comprehensiveness of our study, we employed five graduates with special knowledge to do human evaluation in MMEVALPRO. The designed web ui of human evaluation is shown in figure 15.

F Comparison of Different Models

We list three cases comparing the output of GPT-4o (OpenAI, 2024), GPT-4o (non-vision) (OpenAI, 2024), InternVL-Chat (Chen et al., 2023b) and LLaVA-1.5 (Liu et al., 2023b). We observed that, while both GPT-4o (non-vision) and LLaVA-1.5 could answer all the origin questions correctly, they struggled with most perception and knowledge questions. This highlights a Type-I error in current evaluation benchmarks: correctly answering a question does not necessarily indicate genuine understanding by the model. On the other hand, more advanced models like GPT-4o and InternVL-Chat demonstrate higher consistency in answering different types of questions.


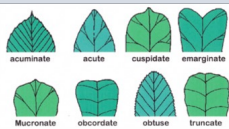
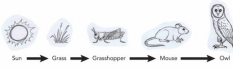
	<p>Original Question</p> <p>Is Dark Orchid the maximum? Options: A. Yes B. No Answer: B</p>	<p>Perception Question</p> <p>How many colors are there in total? Options: A. 5 B. 6 C.7 D.8 Answer: C</p>	<p>Knowledge Question</p> <p>How to determine the maximum here? Options: A. The color which has the biggest area. B. The darkest color. C. The lightest color. Answer: A</p>	<u>1</u>
<p>GPT4-o: B/C/A GPT4-o (non-vision): B/A/A InternVL-Chat: B/D/A LLaVA-1.5-7B: B/A/B</p>				
	<p>Original Question</p> <p>Which leaf has the most veins? Options: A. Acuminate B. Truncate C. Mucronate D. Acute Answer: A</p>	<p>Perception Question</p> <p>How many leaves are there in total? Options: A. 5 B. 6 C.7 D.8 Answer: D</p>	<p>Knowledge Question</p> <p>What are the number of leaf veins for 'Truncate' and 'Obcordate'? Options: A. 6 and 7 B. 5 and 6 C. 7 and 6 D. 7 and 7 Answer: D</p>	<u>2</u>
<p>GPT4-o: A/D/C GPT4-o (non-vision): A/A/A InternVL-Chat: B/D/A LLaVA-1.5-7B: A/A/A</p>				
	<p>Original Question</p> <p>What would be impacted by an increase in owls? Options: A. Sun B. Grass C. Grasshoppers D. Mouse Answer: D</p>	<p>Perception Question</p> <p>What creatures are next to mouse? Options: A. Sun and Grass B. Grasshopper and Owl C. Mouse and Grass D. Sun and Owl Answer: B</p>	<p>Knowledge Question</p> <p>What would happen if the number of owls increase? Options: A. The mouse would increase B. The grass would increase C. The mouse would decrease D. The grasshopper would decrease Answer: C</p>	<u>3</u>
<p>GPT4-o: D/B/C GPT4-o (non-vision): D/A/C InternVL-Chat: D/B/C LLaVA-1.5-7B: D/A/B</p>				

Figure 16: Cases of different models' comparison for MMEVALPRO.