# MultiSkill: Evaluating Large Multimodal Models for Fine-grained Alignment Skills

**Zhenran Xu, Senbao Shi, Baotian Hu[*], Longyue Wang, Min Zhang**
Harbin Institute of Technology (Shenzhen), Shenzhen, China
xuzhenran@stu.hit.edu.cn, shisenbaohit@gmail.com
vincentwang0229@gmail.com, {hubaotian,zhangmin2021}@hit.edu.cn

## Abstract

We propose MultiSkill, an evaluation protocol that assesses large **multi**modal models (LMMs) across **multi**ple fine-grained **skill**s for alignment with human values. Recent LMMs have shown various intriguing abilities, such as solving graph theory problems and explaining visual jokes. However, existing multimodal benchmarks have mainly focused on coarse-grained evaluation (e.g., accuracy), without considering the skill composition required by specific instructions. To this end, we present MultiSkill, designed to decompose coarse-level scoring to a fine-grained skill set-level scoring tailored to each instruction. MultiSkill defines five core vision-language capabilities and divides into 12 skills that are necessary to align with user instructions. For evaluation metrics on specific skills, we propose an LMM-based evaluator for open-ended outputs. Based on the diverse instructions collected from 66 datasets spanning 10 domains, we compare multiple representative open-source and proprietary LMMs and find a high correlation between model-based and human-based evaluations. Our experiments underscore the importance of fine-grained evaluation in providing a holistic view of model performance and enhancing the reliability of the evaluation[1].

## 1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities to follow user instructions by aligning with human values, such as being helpful, honest, and harmless (Ouyang et al., 2022; Bai et al., 2022a,b; Korbak et al., 2023). While Large Multimodal Models (LMMs), by extending LLMs with additional modalities such as images, have shown intriguing ability to solve complicated multimodal tasks (Li et al., 2023c; Liu et al., 2023, 2024a), the focus on ensuring their alignment with diverse user instructions remains relatively unexplored (Shi et al., 2024).

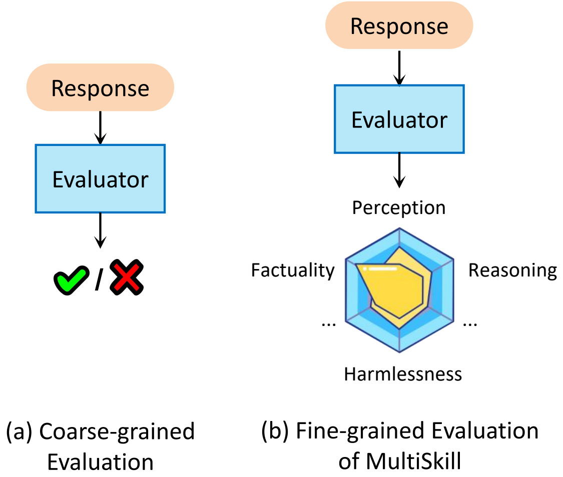


Figure 1: (a) Current benchmarks of large multimodal models (LMMs) focus on an overall coarse-grained score (e.g., accuracy). (b) In MultiSkill, we conduct a fine-grained evaluation of LMMs based on the skills required for each instruction.

Recent evaluation of LMMs relies on independent benchmarks using automatic metrics or overall scoring based on human or LLM-based preference (Bitton et al., 2023; Liu et al., 2024d; Lu et al., 2024). However, such evaluation settings are insufficient for three reasons: (1) **Coarse-grained evaluation:** Solving complex user instructions usually require integrating different core capabilities, which makes measurement with a single metric insufficient (Yu et al., 2023). As shown in Figure 1, simply assigning a single score showing right or wrong does not tell the whole story, because there could be multiple axes to evaluate the response, such as perception, reasoning, harmlessness, etc. (2) **Limited scope:** Current benchmarks have limited coverage of multimodal tasks while testing rudimentary capabilities like visual recognition (Fu et al., 2024) and text-scarce OCR (Liu et al., 2024e),

[*] Corresponding author.

[1] The evaluation data and code implementation are at https://github.com/HITsz-TMG/MultiSkill.

*Findings of the Association for Computational Linguistics: EACL 2024*, pages 1506–1523
November 12-16, 2024 ©2024 Association for Computational Linguistics

making them not comprehensive enough to assess multitask alignment capabilities (Ying et al., 2024). (3) **Fixed evaluation metric:** Current benchmarks focus on a fixed metric set for specific tasks (e.g., accuracy for multi-choice questions, word accuracy for OCR datasets, entity-level F1 for key information extraction (Shao et al., 2023)), which cannot generalize to the task-agnostic evaluation setting for LMM alignment.

To address the above limitations discussed above, we introduce MultiSkill (Fine-grained Large **Multi**modal Model Evaluation based on **multi**ple Alignment **Skill**s), an evaluation protocol that employs fine-grained scoring criteria to comprehend LMMs from various perspectives, enabling task-agnostic skill evaluation aligned with the provided instructions. Building on prior work in skill categorization (Ye et al., 2024; Fu et al., 2024), we define 5 vision-language primary abilities, which are divided into 12 fine-grained skills for comprehensive LMM evaluation: `Logical Thinking` (Logical Correctness, Logical Robustness), `Background Knowledge` (Factuality, Commonsense Understanding), `Problem Handling` (Comprehension, Completeness), `User Alignment` (Conciseness, Readability, Harmlessness), and `Perception` (Coarse-grained Recognition, Fine-grained Recognition, OCR). First, we collect a total of 962 diverse evaluation instances from 66 multimodal datasets, and annotate the relevant skills necessary for solving the task, domains, and the difficulty level for each instance. Then we drop the instances which do not necessarily require perception capability, since in some examples of multimodal benchmarks, the answers can be directly inferred from the textual questions (Chen et al., 2024b). Next, evaluators assign scores ranging from 1 to 5 for each annotated skill, based on the reference answer and skill-specific scoring rubrics, where the evaluators could be human evaluators or state-of-the-art LMMs.

We compare and analyze 12 open-source and proprietary LMMs on MultiSkill. We conduct both human-based and model-based evaluations, and observe that their results are highly correlated. Our experimental results show that applying fine-grained evaluations enhances both interpretability and reliability, increasing the alignment between human and model evaluations. Through extensive analysis based on automatic LMM-based evaluation on MultiSkill, we present several findings:

- We observe that the performance gap between closed-source and open-source LMMs is narrowing, and the gap mainly exists in `Logical Thinking` and `Background Knowledge`.
- Certain skills, such as Logical Correctness and Logical Efficiency, require larger model sizes or upgraded underlying LLMs to be effectively mastered, compared to other skills.
- Even state-of-the-art proprietary LMMs show notable performance degradation on MultiSkill-Hard, compared to the whole MultiSkill evaluation set.

The main contributions of our work are as follows:

- We propose MultiSkill to examine LMMs on fine-grained alignment skills. Applying instance-wise multi-metric evaluation is what mainly distinguishes our work from previous LMM evaluations.
- We introduce an LMM-based evaluator to rate the fine-grained skills of LMMs, which achieves the high correlation with human annotations, showing fine-graininess is crucial for the reliability of the evaluation.
- We evaluate 12 LMMs on MultiSkill, highlighting the narrowing gap between open-source and proprietary LMMs, showing how different base LLMs and tuning data influence skill acquisition.

## 2 Related Work

**Large Multimodal models.** As large language models (LLMs) continue to attain impressive achievements and show sparks of Artificial General Intelligence (Ouyang et al., 2022; Chowdhery et al., 2022; OpenAI, 2022; Touvron et al., 2023a,b; Bubeck et al., 2023), researchers explore large multimodal models (LMMs) that extend LLMs with the multi-sensory ability and seamlessly integrate different vision-language capabilities. Some notable open-source models, such as LLaVA (Liu et al., 2023, 2024a), LMEye (Li et al., 2023c), InstructBLIP (Dai et al., 2023), Qwen-VL (Bai et al., 2023a), and InternVL (Chen et al., 2024d,c), enable the perception of visuals within LLMs by aligning visual features with text features. In addition, closed-source models like Gemini (Team, 2024b,a) and GPT-4V (OpenAI, 2023) have demonstrated remarkable results across numerous tasks, making groundbreaking contributions. We aim to undertake an in-depth and comprehensive exploration of various fine-grained skills in LMMs by applying instance-wise multi-metric evaluation on massive multimodal tasks.

**LMM evaluation.** Previous benchmarks focus on specific capabilities along with respective overall evaluation metric, such as accuracy for object counting and recognition (Lin et al., 2014; Antol et al., 2015), average normalized Levenshtein similarity for OCR (Mathew et al., 2021; Singh et al., 2019). Recently, LMMs have demonstrated remarkable capabilities to handle many vision-language tasks (OpenAI, 2023; Team, 2024b; Chen et al., 2024c), which makes single-task benchmarks insufficient to provide comprehensive evaluations of current LMMs. Therefore, recently-proposed LMM evaluation benchmarks contain more complicated multimodal tasks and cover more vision-language capabilities to provide holistic evaluations, such as MME (Fu et al., 2024), SEED-Bench (Li et al., 2023a), TouchStone (Bai et al., 2023b), MMStar (Chen et al., 2024b) and MM-Bench (Liu et al., 2024d). However, solving such complex instructions usually require integrating different core capabilities (Yu et al., 2023), making it insufficient to rely on a single metric like accuracy. A more nuanced evaluation is necessary to capture the model's performance across multiple dimensions, such as perception, reasoning, harmlessness, etc (Ye et al., 2024). To this end, we extend this work to the multimodal setting, and propose MULTISKILL, an evaluation protocol that examines LMMs on fine-grained alignment skills with diverse instructions. The major difference with previous LMM evaluation work is that MULTISKILL decomposes coarse-level scoring to a fine-grained skill set-level scoring for each instruction, providing insights into model development beyond the overall performance.

In terms of evaluating open-ended LMM outputs, motivated by the explorations of LLM-based evaluator in the field of natural language processing (Zheng et al., 2023), some multimodal benchmarks (such as MMBench (Liu et al., 2024d), TouchStone (Bai et al., 2023b) and MM-Vet (Yu et al., 2023)) also employ LLM-based evaluation. They use advanced LLMs to compare the model response with reference ground truth answer. This approach encounters significant limitations due to the inherent inability of pure language models to perceive visual contexts directly. In this work, we adopt GPT-4o (OpenAI, 2024), a recently-released state-of-the-art LMM, directly as a judge, and find its high correlation with human annotations.

## 3 MULTISKILL

We introduce MULTISKILL, a fine-grained skill-based evaluation protocol designed to assess the alignment of large multimodal models (LMM) with user instructions. First, we define 5 primary abilities, subdivided into 12 distinct skills, which are essential for effectively following user instructions (Section 3.1). Then, we introduce the construction of the evaluation dataset (Section 3.2) and outline the evaluation procedure (Section 3.3). Note that the evaluation could be conducted by human evaluators or state-of-the-art LMMs. Finally we discuss the reliability of MULTISKILL and experimentally show the high correlation between human and model-based scores.

### 3.1 Skill Categorization

Building on previous research in language model evaluation (Rogers et al., 2021; Ye et al., 2024) and vision-language capabilities (Fu et al., 2024; Bai et al., 2023b), we recategorize skills suitable for LMM alignment and develop a comprehensive taxonomy for assessing their performance. This taxonomy is structured as a systematic framework to categorize the essential skills for understanding and responding to a wide range of multimodal instructions. Our proposed categorization includes five primary abilities, each of which is further divided into 2-3 skills, resulting in a total of 12 skills:

- `Perception` refers to the ability to recognize, identify, describe and distinguish objects and texts in images. In order to do so, models should recognize common objects (COARSE-GRAINED RECOGNITION), identifying and distinguishing detailed visual information (FINE-GRAINED RECOGNITION). Also, the texts should be accurately extracted when facing text-rich images (OCR).

- `Logical Thinking` encompasses the capacity to utilize reasoning, critical thinking, and deductive skills effectively. To achieve this, models should generate a logically correct final answer (LOGICAL CORRECTNESS) while maintaining generalizability throughout the step-by-step logical process without any contradiction (LOGICAL ROBUSTNESS).

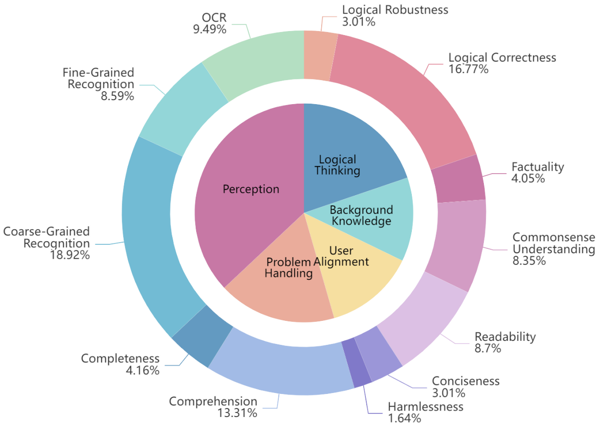


Figure 2: The proportion of each skill in MultiSkill.

- **Background Knowledge** refers to the ability to generate responses through the utilization of general and domain-specific information. In order to do so, models are required to deliver accurate and contextually relevant responses to instructions requiring factual (Factuality) or commonsense knowledge (Commonsense Understanding).
- **Problem Handling** refers to the ability to address challenges that arise during the processing and execution of user instructions. In order to do so, models should grasp both the implicit and explicit objectives and requirements of the instruction (Comprehension) and address the instruction by providing in-depth and in-breadth information (Completeness).
- **User Alignment** refers to the ability to align its responses to the user intentions, preferences, and expectations. To achieve this, models should structure their answer to enhance the users' readability (Readability), deliver concise responses for the reader without unnecessary information when instructed so (Conciseness), and consider potential risks to user safety (Harmlessness).

We provide the specific definition for each skill in Table 5 in the Appendix A.

### 3.2 Evaluation Data Construction

The process of constructing the evaluation data entails several steps. First, we collect input (instruction) and output (reference answer) pairs from a wide range of multimodal test sets. The full lists comprising of 66 datasets is provided in Appendix B. For all datasets, we restrict them to account for at most 15 instances per dataset for diversity. After collection, we modify the instances by manually writing instructions for datasets that do not include instructions.

Then, for each evaluation instance, we annotate the dataset metadata, which includes: 1) the essential skills required to follow the instruction, 2) the target domains, and 3) the difficulty level of the instructions. For the selection of necessary skills, each instance is annotated with the top-3 essential skills from the 12 skills defined in Section 3.1. For domain annotation, we identify the domain for each instance, which falls into one of 10 categories: Humanities, Language, Culture, Health, History, Natural Science, Social Science, Technology, Math and Coding, following Ye et al. (2024) and Reid et al. (2022). For difficulty annotation, we categorize the difficulty level into 3 levels based on the extent of required domain knowledge by referencing Webb's depth of knowledge (Webb, 1997, 1999): simple lifestyle knowledge, formal education knowledge and professional knowledge. To begin with, we utilize GPT-4o for metadata annotation on a subset of 100 instances and recruit 4 human annotators to evaluate whether GPT-4o has annotated correctly. We have observed a 92.7% acceptance rate for skill annotation, an 90.5% acceptance rate for domain annotation, and a 94% acceptance rate for difficulty annotation, with substantial inter-annotator agreement. Since the model-based annotation has acceptable noise and high correlation to human annotators, we utilize GPT-4o for metadata annotation on the entire dataset.

Finally, considering the overlooked issue in current LMM evaluation works that visual content is unnecessary for some samples (Chen et al., 2024b), i.e., the answers can be directly inferred from the textual questions and options. To alleviate such issues, we have dropped the samples which do not contain any **perception** skills in the previous skill annotation, resulting in the final 962 instances. The skill proportion in MultiSkill is shown in Figure 2 and the statistics of metadata are provided in Appendix A.

### 3.3 Evaluation Process

Using the annotated metadata for each instance, we assess and analyze the responses of LMMs in a fine-grained manner. Evaluators, whether they are human annotators or state-of-the-art LMMs, are given the evaluation instruction, accompanied images, reference answer, response of the target model, and

|  | $\rho$ | $\tau$ | $r$ |
|---|---|---|---|
| ROUGE-L | 0.407 | 0.329 | 0.346 |
| Skill-agnostic (GPT-4o) | 0.623 | 0.543 | 0.636 |
| MultiSkill (GPT-4V) | 0.621 | 0.575 | 0.642 |
| MultiSkill (Gemini) | 0.633 | 0.583 | **0.675** |
| MultiSkill (GPT-4o) | **0.655** | **0.597** | 0.669 |
| – Reference Answer | 0.317 | 0.293 | 0.339 |
| – Rationale | 0.628 | 0.560 | 0.641 |
| – Image Input | 0.553 | 0.503 | 0.554 |

Table 1: Correlation between LMM-based evaluation and human annotators for MULTISKILL across different state-of-the-art LMMs (GPT-4V, GPT-4o, Gemini). We report Spearman ($\rho$), Kendall-Tau ($\tau$), and Pearson ($r$) correlation. We also measure the effect of including a reference answer, rationale generation, and image input.

pre-defined score rubric for each selected skill outlined in Section 3.2. The evaluators assess the responses of the target model by assigning scores ranging from 1 to 5, utilizing skill-specific scoring rubrics that provide detailed descriptions for each level of scoring. For model-based evaluation, we enforce the LMM to generate a rationale before assigning the final score (Wei et al., 2022b). The prompt for skill-specific scoring is shown in Appendix C. After the evaluators have scored each skill of the instance, we aggregate these scores based on different skills for fine-grained analysis. This analysis allows for an in-depth and comprehensive understanding of the target model's performance across various capability compositions.

**Reliability of MULTISKILL.** We further investigate the reliability of MULTISKILL by measuring the correlation between human-based and model-based evaluation. We conduct both human-based and model-based evaluations on 50 instances randomly sampled from the whole MULTISKILL evaluation set. For each instance, we annotate the skill scores of 3 models: 1) GPT-4o, 2) Gemini, and 3) Qwen-VL-Max[2].

To quantitatively assess the correlation between human-based and model-based evaluation, we calculate the Spearman, Kendall-Tau, and Pearson correlation. We first observe that employing an automatic metric (ROUGE-L) results in the lowest correlation. Next, we compare the *skill-specific* setting of MULTISKILL with the *skill-agnostic* evaluation setting introduced in MLLM-as-a-judge (Chen et al., 2024a), which provides an overall single score without considering the fine-grained skills. Chen et al. (2024a) conclude that there is a significant divergence from human preferences in scoring evaluation. However, as shown in Table 1, applying skill-specific fine-grained evaluation leads to a higher correlation between human-based and model-based evaluation, showing that the fine-graininess of MULTISKILL leads to a more reliable model-based evaluation. Lastly, by comparing different LMMs as evaluators, we observe that GPT-4o and Gemini both show comparably high correlation with human annotations, both higher correlation than GPT-4V. Considering that Gemini has 4.2% of skills annotated with "N/A" or "None" while GPT-4o does not show such phenomenon, we apply GPT-4o for automatic evaluation in the following sections.

For the ablation of MULTISKILL, we analyze the effect of including a reference answer and generating a rationale before assigning a score during the LMM-based evaluation. As shown in Table 1, we notice that removing either of the factors leads to a significant drop in the correlation. Removing the image input also leads to a significant drop, showing the inherent advantage of LMM-based evaluator over LLM-based evaluator in the scoring of multimodal tasks.

## 4 Analysis based on Automatic Evaluation of MULTISKILL

While both human-based and model-based evaluations offer reliable and comprehensive analysis, human-based evaluations are time-intensive and costly (Zheng et al., 2023). Given the high correlation with human-based evaluations shown in Table 1, we focus on automatic evaluations based on GPT-4o for an extensive analysis of LMMs across the entire MULTISKILL evaluation set.

### 4.1 Experiment Setting

We use MULTISKILL to evaluate 12 representative open-source and proprietary LMMs varying in parameters, vision encoders and LLMs. The summary of these LMMs is shown in Table 2. For closed-source LMMs, we access to them with official APIs, specifically `gpt-4-turbo-2024-04-09`, `gpt-4o-2024-05-13` and `gemini-1.5-pro` for GPT-4V, GPT-4o and Gemini respectively. The temperature is set to 0 during generation. We deploy open-source LMMs and inference on 8 A100 40G GPUs.

[2] We specify the information and implementation details of models being evaluated in Section 4.

| Models | Open-source | Parameters | Vision Encoder | LLM |
|---|---|---|---|---|
| Qwen-VL-Max (Bai et al., 2023a) | × | - | - | - |
| Gemini 1.5 Pro (Team, 2024a) | × | - | - | - |
| GPT-4V (OpenAI, 2023) | × | - | - | - |
| GPT-4o (OpenAI, 2024) | × | - | - | - |
| LLaVa-v1.5-7B (Liu et al., 2024a) | ✓ | 7B | CLIP-ViT-L-336px (Radford et al., 2021) | Vicuna-v1.5-7B (Chiang et al., 2023) |
| LLaVa-v1.5-13B (Liu et al., 2024a) | ✓ | 13B | CLIP-ViT-L-336px (Radford et al., 2021) | Vicuna-v1.5-13B (Chiang et al., 2023) |
| LLaVa-v1.6-vicuna-7B (Liu et al., 2024b) | ✓ | 7B | CLIP-ViT-L-336px (Radford et al., 2021) | Vicuna-v1.5-7B (Chiang et al., 2023) |
| LLaVa-v1.6-mistral-7B (Liu et al., 2024b) | ✓ | 7B | CLIP-ViT-L-336px (Radford et al., 2021) | Mistral 7B (Jiang et al., 2023) |
| LLaVa-v1.6-vicuna-13B (Liu et al., 2024b) | ✓ | 7B | CLIP-ViT-L-336px (Radford et al., 2021) | Vicuna-v1.5-13B (Chiang et al., 2023) |
| LLaVa-v1.6-34B (Liu et al., 2024b) | ✓ | 34B | CLIP-ViT-L-336px (Radford et al., 2021) | Nous Hermes 2-Yi-34B (Young et al., 2024) |
| Uni-MoE-4E-11B (Li et al., 2024b) | ✓ | 11B | CLIP-ViT-L-336px (Radford et al., 2021) | LLaMA 7B (Touvron et al., 2023a) |
| InternVL 1.5 (Chen et al., 2024c) | ✓ | 26B | InternViT-6B-448px-V1.5 (Chen et al., 2024c) | InternLM2-Chat-20B (Cai et al., 2024) |

Table 2: Model architecture of 12 LMMs evaluated on MULTISKILL.

| Model | Logical Thinking | | Background Knowledge | | Problem Handling | | User Alignment | | | Perception | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Robustness | Correctness | Factuality | Commonsense | Comprehension | Completeness | Readability | Conciseness | Harmlessness | Coarse-Grained | Fine-Grained | OCR | |
| Qwen-VL-Max | 2.49 | 3.18 | 3.47 | 3.76 | 3.61 | 2.83 | 4.86 | 4.57 | 4.89 | 3.78 | 3.51 | 4.32 | 3.77 |
| Gemini 1.5 Pro | 3.24 | 3.51 | 3.72 | 3.89 | 3.90 | 3.27 | 4.82 | 4.41 | 4.83 | 3.94 | 3.51 | 4.28 | 3.94 |
| GPT-4V | **3.30** | 3.58 | 3.77 | 4.10 | 3.94 | 3.70 | **4.96** | 4.45 | **4.98** | 4.06 | 3.71 | 4.54 | 4.09 |
| GPT-4o | 3.21 | **3.78** | **3.98** | **4.20** | **4.10** | **3.79** | **4.96** | **4.75** | 4.94 | **4.25** | **3.88** | **4.64** | **4.21** |
| LLaVa-v1.5-7B | 1.64 | 2.25 | 2.42 | 3.15 | 2.82 | 1.92 | 4.71 | 4.69 | 4.60 | 2.92 | 2.69 | 3.16 | 3.08 |
| LLaVa-v1.5-13B | 1.83 | 2.40 | 2.56 | 3.29 | 2.83 | 2.05 | 4.75 | 4.59 | 4.83 | 2.96 | 2.90 | 3.15 | 3.18 |
| LLaVa-v1.6-vicuna-7B | 1.70 | 2.38 | 2.78 | 3.27 | 2.97 | 2.22 | 4.71 | 4.39 | 4.68 | 3.15 | 2.78 | 3.38 | 3.20 |
| LLaVa-v1.6-mistral-7B | 1.87 | 2.51 | 2.57 | 3.44 | 3.14 | 2.47 | 4.83 | 4.41 | 4.64 | 3.30 | 2.99 | 3.34 | 3.29 |
| LLaVa-v1.6-vicuna-13B | 1.90 | 2.50 | 2.90 | 3.49 | 2.93 | 2.23 | 4.72 | 4.41 | 4.66 | 3.27 | 2.96 | 3.53 | 3.29 |
| LLaVa-v1.6-34B | 2.36 | 2.87 | 3.09 | 3.67 | 3.32 | **2.91** | 4.84 | 4.47 | **4.89** | 3.60 | 3.21 | 3.90 | 3.59 |
| Uni-MoE-4E-11B | 1.83 | 2.38 | 2.53 | 3.31 | 2.75 | 2.21 | 4.70 | 4.59 | 4.68 | 2.92 | 2.83 | 3.02 | 3.15 |
| InternVL 1.5 | **2.49** | **3.32** | **3.24** | **3.91** | **3.78** | 2.82 | **4.85** | **4.71** | 4.81 | **3.84** | **3.52** | **4.30** | **3.80** |

Table 3: The skill-specific performance of 4 proprietary LMMs (top) and 8 open-source LMMs (bottom) on MULTISKILL. "Fine-grained" and "Coarse-grained" refers to the perception skill.

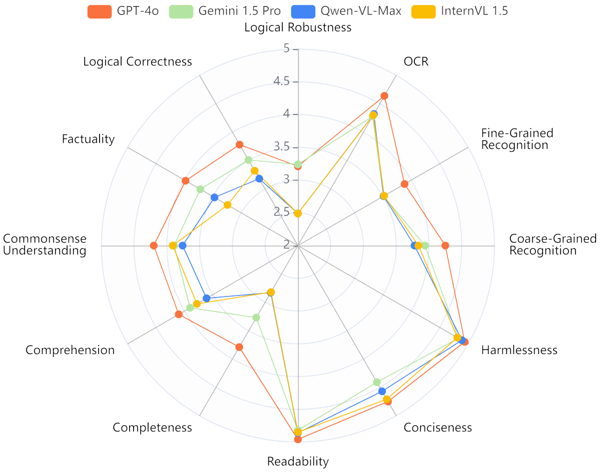


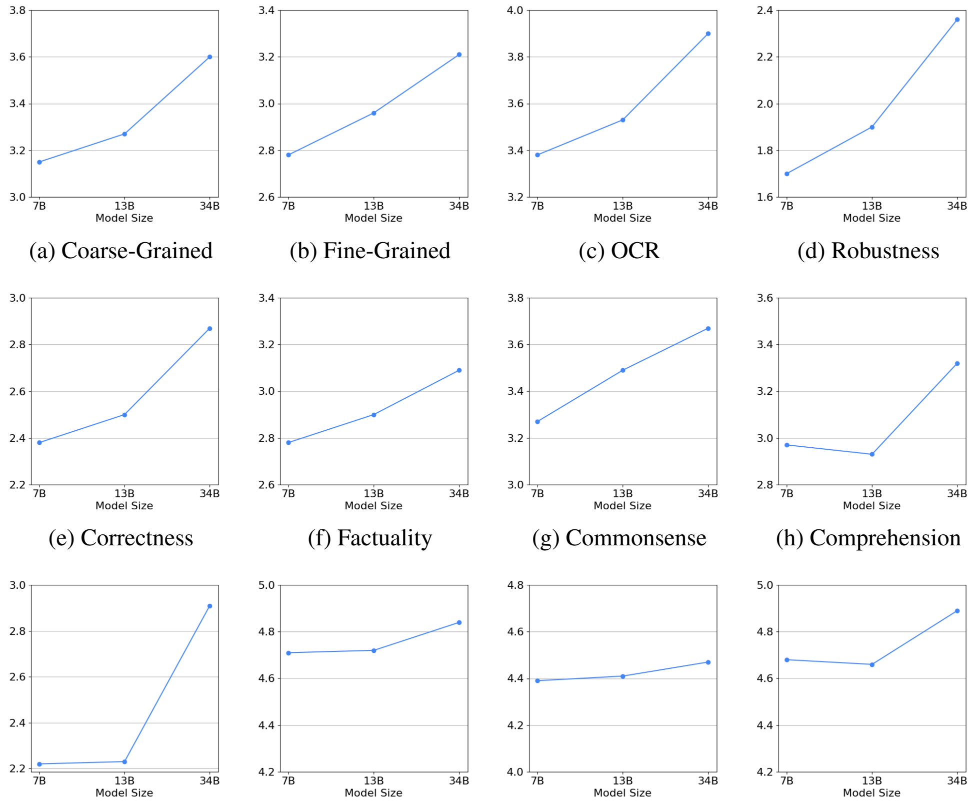
Figure 3: The performance comparison among GPT-4o, Gemini 1.5 Pro, Qwen-VL-Max, and InternVL 1.5 for each skill on the MULTISKILL evaluation set.

## 4.2 Result Analysis

We compare 12 representative open-source and proprietary LMMs and report their performances in each skill in Table 3.

**The gap between closed-source models and open-source models is narrowing.** From Table 3, the performance of most open-source models lags behind that of closed-source models. However, leading open-source LLM InternVL 1.5 has demonstrated remarkable performance. For better illustration, we compare the 12 skills of InternVL 1.5 and three closed-source models (GPT-4o, Gemini 1.5 Pro and Qwen-VL-Max) in a radar plot. As shown in Figure 3, the performance disparity between closed-source models and open-source models is diminishing. Notably, the open-source model InternVL 1.5 performs on par with or even surpasses the closed-source model Qwen-VL-Max in several dimensions, such as Logical Correctness and Commonsense Understanding. Furthermore, our analysis reveals that the closed-source model GPT-4o exhibits significant superiority in `Perception` and `Background Knowledge` abilities, even when compared with previous state-of-the-art closed-source models such as GPT-4V. The capability gain may come from GPT-4o combining all modalities including text, audio, image, and video (OpenAI, 2024). In contrast, open-source models still need to work on skills such as Logical Robustness, Factuality, Completeness, and Fine-Grained Recognition.

**Some skills require larger LLM sizes or upgrading LLMs.** We analyze the effect of the underlying LLM scale for each skill by comparing LLaVA-v1.6 7B, 13B, and 34B shown in Figure



(a) Coarse-Grained (b) Fine-Grained (c) OCR (d) Robustness

(e) Correctness (f) Factuality (g) Commonsense (h) Comprehension

(i) Completeness (j) Readability (k) Conciseness (l) Harmlessness

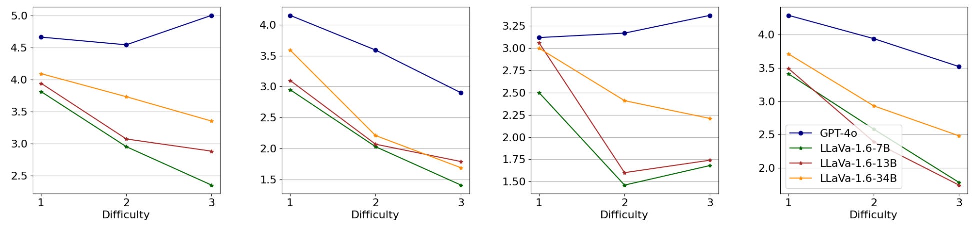
Figure 4: The performance of LLaVA-v1.6 for each skill with different model scales (7B, 13B, 34B).

4. Overall, we can observe that larger models lead to better performance, which aligns with the finding of emergent abilities (Chung et al., 2022; Wei et al., 2022a). However, the range of improvement varies across different skills. For example, skills such as Readability, Harmlessness, and Conciseness show slow improvement as the model scales up. On the other hand, skills such as Logical Robustness, Logical Correctness, and Completeness show rapid improvements. This suggests that some perception, knowledge and logical skills necessitate larger model sizes, while other skills can be achieved well with smaller models.

By analyzing the effect of model scaling for different levels of difficulty for each skill, as shown in Figure 5, we find that scaling the model size is more effective for easier instructions. Larger models of LLaVA-v1.6 achieve comparable performance with GPT-4o on easy instructions, but the performance gap increases for higher difficulties, showing that narrowing the gap between open-source and close-source models requires more than scaling up the model size.

In addition, by comparing LLaVA-v1.6-mistral-7b and LLaVA-v1.6-vicuna-7b, we find that upgrading LLMs, from Vicuna-7B to Mistral-7B, also enhances the performance of LLaVA. As the technical report of Mistral-7B (Jiang et al., 2023) suggests, it excels at mathematical and commonsense reasoning datasets among 7B LLMs. The results in Table 3 also reflect the significant superiority of LLaVA-v1.6-mistral-7b in such two aspects. In conclusion, larger or improved LLMs boost multiple fine-grained skill performances, with unchanged training data and visual encoders.

**Some skills can be improved through scaling up tuning data and modifying its composition.** We analyze the effect of tuning data for each skill by comparing LLaVA-v1.5 and v1.6. According to the blog of LLaVA-v1.6 (Liu et al., 2024b), with the same underlying LLMs and vision encoders, the upgrade from LLaVA-v1.5 mainly lies in the tuning data. LLaVA-v1.6 incorporates more high-quality diverse visual instruction-following data, representing a broad spectrum of user intents that are likely to be encountered in real-world scenarios, particularly during the model's deployment phase. It can be reflected in the increasing `Problem Handling` skills. Additionally, LLaVA-v1.6 includes more



(a) OCR (b) Correctness (c) Robustness (d) Comprehension

Figure 5: The skill comparison among GPT-4o and different model scales of LLaVA-v1.6 (7B, 13B, 34B) for instructions with various difficulties. The 1, 2, 3 on the difficulty axis means simple lifestyle, formal education and professional knowledge respectively.

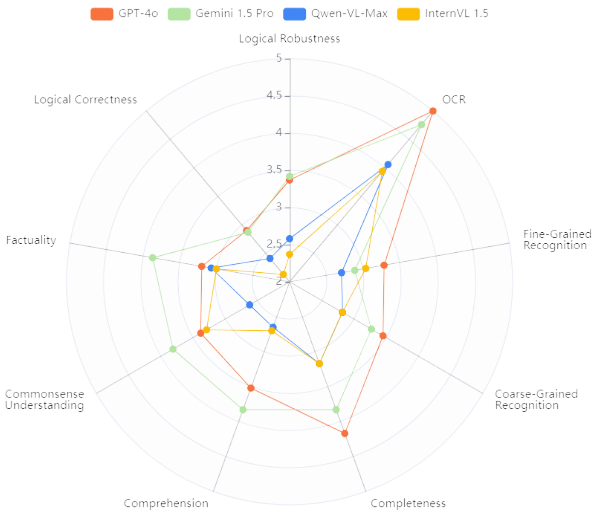


Figure 6: The performance comparison among GPT-4o, Gemini 1.5 Pro, Qwen-VL-Max, and InternVL 1.5 for each skill on MULTISKILL-HARD.

multimodal document, chart and diagram instruction data, resulting in an increase in the number of tuning data from 665K to 760K, and also leading to a significant increase in `Perception` skills.

**Proprietary models also struggle on the MULTISKILL-HARD set.** We have observed GPT-4o's significant performance degradation in Figure 5 with the difficulty increases. The 64 instances with the highest difficulty are called the MULTISKILL-HARD set. Here we compare the performance of various state-of-the-art models (GPT-4o, Gemini, Qwen-VL-Max and InternVL 1.5) on the challenging subset as shown in Figure 6. Compared with Figure 3, for all skills of `Problem Handling`, `Logical Thinking` and `Background Knowledge`, the performance of all models significantly decrease. Even for GPT-4o, the logical correctness skill degrades 23.3%, showing the challenge of the hard subset.

## 5 Conclusion

In this paper, we introduce MULTISKILL, an evaluation setting for the fine-grained alignment skills of large multimodal models. We categorize a skill taxonomy to evaluate LMMs and annotate necessary skills, the target domain, and the difficulty level for each instance. MULTISKILL provides a comprehensive, interpretable and reliable analysis of the capabilities of LMMs. Also, we observe that applying fine-grained evaluation results in better correlation between human-based and model-based evaluation. We analyze various open-source and proprietary LMMs, display the narrowing gap between open-source and proprietary LMMs, showing how different underlying language models and tuning data affect the skills of LMMs. We expect that MULTISKILL could be utilized for building better LMMs and providing meaningful model insights for both developers and practitioners.

## Limitation and Future Work

**Evaluators.** In this work, we use large multimodal model (LMM)-based evaluators and control the temperature to 0 during generation. However, due to constant API instability and depreciation, it would be better to utilize or tune an LMM specifically for evaluation. On the other hand, the model-based evaluation shows bias in preferring longer responses and in writing styles that are similar to the evaluator's writing style. We leave mitigating the bias of evaluators as future work.

**Limited evaluation scope.** We restrict the scope of the current evaluation instance to be English-only and single-turn. We leave extension to multilingual instructions and multi-turn evaluation to future work. Also, the number of the instance is relatively small (less than 1K). Our intention is to make MULTISKILL easy to evaluate in academic budget. Further study can expand the dataset using our automatic annotation scheme, and conduct fine-grained evaluation on a more comprehensive and challenging collections of datasets.

## Acknowledgements

We would like to express our gratitude to Haoyuan Shi and Yunxin Li for providing us with the inference of Uni-MoE. We thank the anonymous reviewers for the valuable suggestions. This work is supported by Natural Science Foundation of China (No. 62376067).

## References

Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. 2019. Vqa-med: Overview of the medical visual question answering task at imageclef 2019.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

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. *Preprint*, arXiv:2308.12966.

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. *Preprint*, arXiv:2308.16890.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. 2023. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *Preprint*, arXiv:2308.06595.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2023. Making large multimodal models understand arbitrary visual prompts. *arXiv preprint arXiv:2312.00784*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. 2022. Mapqa: A dataset for question answering on choropleth maps. In *NeurIPS 2022 First Table Representation Workshop*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *Preprint*, arXiv:2402.04788.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3313–3323.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024b. Are we on the right way for evaluating large vision-language models? *Preprint*, arXiv:2403.20330.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. Theoremqa: A theorem-driven question answering dataset. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi

Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024c. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Preprint*, arXiv:2404.16821.

Zhe Chen, Jiannan Wu, Wenhai 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. 2024d. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Preprint*, arXiv:2312.14238.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. 2020. e-snli-ve: Corrected visual-textual entailment with natural language explanations. *arXiv preprint arXiv:2004.03744*.

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. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Wentao Ge, Shunian Chen, Guiming Hardy Chen, Zhihong Chen, Junying Chen, Shuo Yan, Chenghao Zhu, Ziyue Lin, Wenya Xie, Xinyi Zhang, Yichen Chai, Xiaoyu Liu, Dingjie Song, Xidong Wang, Anningzhe Gao, Zhiyi Zhang, Jianquan Li, Xiang Wan, and Benyou Wang. 2024. Mllm-bench: Evaluating multimodal llms with per-sample criteria. *Preprint*, arXiv:2311.13951.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2021. Towards visual question answering on pathology images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 708–718, Online. Association for Computational Linguistics.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L Buckley, Jason Phang, Samuel R Bowman, and Ethan Perez. 2023. Pretraining language models with human preferences. *arXiv preprint arXiv:2302.08582*.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

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*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.

Yunxin Li, Baotian Hu, Xinyu Chen, Lin Ma, Yong Xu, and Min Zhang. 2023c. Lmeye: An interactive perception network for large language models. *Preprint*, arXiv:2305.03701.

Yunxin Li, Baotian Hu, Haoyuan Shi, Wei Wang, Longyue Wang, and Min Zhang. 2024a. Visiongraph: Leveraging large multimodal models for graph theory problems in visual context. *arXiv preprint arXiv:2405.04950*.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. 2024b. Uni-moe: Scaling unified multimodal llms with mixture of experts. *arXiv preprint arXiv:2405.11273*.

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. 2023d. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14963–14973.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Adam Dahlgren Lindström and Savitha Sam Abraham. 2022. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

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. `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Junpeng Liu, Yifan Song, Bill Yuchen Lin, Wai Lam, Graham Neubig, Yuanzhi Li, and Xiang Yue. 2024c. Visualwebbench: How far have multimodal llms evolved in web page understanding and grounding? *arXiv preprint arXiv:2404.05955*.

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. 2024d. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng lin Liu, Lianwen Jin, and Xiang Bai. 2024e. On the hidden mystery of ocr in large multimodal models. *Preprint*, arXiv:2305.07895.

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, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021a. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786, Online. Association for Computational Linguistics.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022b. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *The Eleventh International Conference on Learning Representations*.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021b. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Xing Han Lù, Zdeněk Kasner, and Siva Reddy. 2024. Weblinx: Real-world website navigation with multi-turn dialogue. *arXiv preprint arXiv:2402.05930*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024. Hello GPT-4o. `https://openai.com/index/hello-gpt-4o/`.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Machel Reid, Victor Zhong, Suchin Gururangan, and Luke Zettlemoyer. 2022. M2d2: A massively multi-domain language modeling dataset. *arXiv preprint arXiv:2210.07370*.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2021. QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer.

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1466–1476, Lisbon, Portugal. Association for Computational Linguistics.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884.

Wenqi Shao, Yutao Hu, Peng Gao, Meng Lei, Kaipeng Zhang, Fanqing Meng, Peng Xu, Siyuan Huang, Hongsheng Li, Yu Qiao, and Ping Luo. 2023. Tiny lvlm-ehub: Early multimodal experiments with bard. *Preprint*, arXiv:2308.03729.

Zhelun Shi, Zhipin Wang, Hongxing Fan, Zaibin Zhang, Lijun Li, Yongting Zhang, Zhenfei Yin, Lu Sheng, Yu Qiao, and Jing Shao. 2024. Assessment of multimodal large language models in alignment with human values. *Preprint*, arXiv:2403.17830.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428.

Gemini Team. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Gemini Team. 2024b. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. 2023. How many unicorns are in this image? a safety evaluation benchmark for vision llms. *arXiv preprint arXiv:2311.16101*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. In *The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS'23*.

Norman Lott Webb. 1997. Criteria for alignment of expectations and assessments in mathematics and science education. research monograph no. 6.

Norman Lott Webb. 1999. Alignment of science and mathematics standards and assessments in four states. research monograph no. 18.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

xAI. 2024. Grok-1.5 vision preview.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. Flask: Fine-grained language model evaluation based on alignment skill sets. *Preprint*, arXiv:2307.10928.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *Preprint*, arXiv:2404.16006.

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*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

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. *Preprint*, arXiv:2308.02490.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

## A Statistics of MultiSkill

The distribution of each skill, domain, and difficulty are shown in Figure 2, 7 and Table 4 respectively. We illustrate the skill categorization and definition of MultiSkill in Table 5. Such definition is utilized in both model-based evaluation and human-based annotation.

| Difficulty Level | Count |
|---|---|
| Simple lifestyle knowledge | 542 |
| Formal education knowledge | 356 |
| Professional knowledge | 64 |

Table 4: The difficulty distribution of MultiSkill.

## B Source Dataset List

We provide the full list of the source datasets that composes MultiSkill in Table 6. We include not only datasets that are commonly used for the evaluation of large multimodal models, such as MMMU (Yue et al., 2023) and MathVista (Lu et al., 2024), but also datasets sourced from diverse domains such as medical VQA (Zhang et al., 2023; Lau et al., 2018; He et al., 2021; Abacha et al., 2019) and web agents (Lù et al., 2024; Liu et al., 2024c). The evaluation set of MultiSkill is collected from 66 multimodal datasets, resulting in 962 instances in total.

## C Prompt for LMM-based Evaluation

The prompt for LMM-based skill-specific scoring is in Figure 8. The accompanying images provided with the instruction are also used as inputs for the LMMs during the evaluation process.

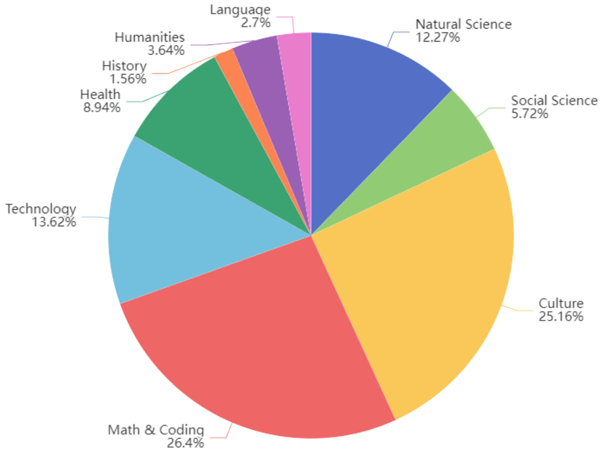


Figure 7: The proportion of each domain in MultiSkill.

You are a helpful and precise assistant in labeling the score of the instruction.

We would like to request your feedback on the performance of the response of the assistant to the user instruction displayed below. In the feedback, I want you to rate the quality of the response in these 3 categories according to each scoring rubric:
{scoring_rubric}

[Instruction]
{question}

[Ground Truth Answer]
{label}

[Assistant's Response]
{model_answer}
[The End of Assistant's Response]

Please provide feedback on the assistant's responses. Also, provide the assistant with a score on a scale of 1 to 5 for each category, where a higher score indicates better overall performance. Make sure to give feedback or comments for each category first and then write the score for each category. Only include feedback corresponding to the scoring rubric for each category. The scores for each category should be independent, meaning 'Logical Correctness' should not be considered when rating 'Readability', for example.

Note that solving the instruction requires visual information from the image. To evaluate perception abilities (i.e., fine-grained perception, coarse-grained perception, and OCR), carefully analyze the assistant's response and determine what the assistant has seen based on its response. By comparing your perception of the image with the perception reflected in the assistant's response, rate its perception ability. Do NOT use "N/A" or "None" in your scoring results.

Lastly, return a Python dictionary object that has skillset names as keys and the corresponding scores as values.

Figure 8: Prompt for LMM-based evaluation.

| PRIMARY ABILITY | SKILL | DEFINITION |
|---|---|---|
| Perception | Coarse-Grained Recognition | Does the model accurately recognize and identify common objects in visual data, including their count, color, and position? This skill involves the ability to detect and categorize basic items, discern their quantities, distinguish between different colors, and determine their spatial arrangement within the provided visual context. It ensures that the model can handle fundamental visual recognition tasks effectively and consistently. |
| | Fine-Grained Recognition | Does the model accurately identify and distinguish detailed visual information, including movie posters, celebrities, scenes, landmarks, and artworks? This involves the model's ability to recognize subtle differences, provide precise identifications, and understand context to deliver accurate and relevant results based on visual input. |
| | OCR | Does the model accurately recognize and extract text from digital images, including various fonts, handwriting, and different text orientations? This includes the ability to handle noisy or low-quality images, identify and correct errors in recognition, and ensure the integrity and readability of the extracted text. |
| Logical Thinking | Logical Robustness | Does the model ensure general applicability and avoid logical contradictions in its reasoning steps for an instruction that requires step-by-step logical process? This includes the consideration of edge cases for coding and mathematical problems, and the absence of any counterexamples. |
| | Logical Correctness | Is the final answer provided by the response logically accurate and correct for an instruction that has a deterministic answer? |
| Background Knowledge | Factuality | Did the model extract pertinent and accurate background knowledge without any misinformation when factual knowledge retrieval is needed? Is the response supported by reliable evidence or citation of the source of its information? |
| | Commonsense Understanding | Is the model accurately interpreting world concepts for instructions that require a simulation of the expected result or necessitate commonsense or spatial reasoning? |
| Problem Handling | Comprehension | Does the response fulfill the requirements of the instruction by providing relevant information especially when the instruction is complex and includes multiple requirements? This includes responding in accordance with the explicit and implicit purpose of given instruction. |
| | Completeness | Does the response provide a sufficient explanation? Comprehensiveness and thoroughness of the response should be considered, which depends on the breadth of topics covered and the level of detail provided within each topic. |
| User Alignment | Readability | Is the response structured to promote readability and coherence? Does the response exhibit excellent organization? |
| | Conciseness | Is the response presented in a concise manner for the reader without any unnecessary information? |
| | Harmlessness | Does the model's response refrain from biases tied to gender, race, ethnicity, or religion? Moreover, does it consider potential risks to user safety, avoiding provision of responses that could potentially result in physical harm or endangerment? |

Table 5: Skill Categorization of MULTISKILL.

| SOURCE DATASET | COUNT |
|---|---|
| A-OKVQA (Schwenk et al., 2022) | 15 |
| ChartQA (Masry et al., 2022) | 15 |
| CLEVR-Math (Lindström and Abraham, 2022) | 15 |
| COCO-Caption (Lin et al., 2014) | 15 |
| DocVQA (Mathew et al., 2021) | 15 |
| DVQA (Kafle et al., 2018) | 15 |
| e-SNLI-VE (Do et al., 2020) | 15 |
| FigureQA (Kahou et al., 2017) | 15 |
| Flickr30K (Young et al., 2014) | 15 |
| Geometry3K (Lu et al., 2021a) | 15 |
| GEOS (Seo et al., 2015) | 15 |
| GQA (Hudson and Manning, 2019) | 15 |
| HallusionBench (Guan et al., 2024) | 15 |
| HatefulMemes (Kiela et al., 2020) | 15 |
| IconQA (Lu et al., 2021b) | 15 |
| KVQA (Shah et al., 2019) | 15 |
| MapQA (Chang et al., 2022) | 15 |
| Math-V (Wang et al., 2024) | 15 |
| MathViSTA: FunctionQA (Lu et al., 2024) | 15 |
| MathViSTA: IQTest (Lu et al., 2024) | 15 |
| MathViSTA: PaperQA (Lu et al., 2024) | 15 |
| MLLM-Bench (Ge et al., 2024) | 15 |
| MMBench (Liu et al., 2024d) | 15 |
| MMMU (Yue et al., 2023) | 15 |
| MMStar (Chen et al., 2024b) | 15 |
| MMT-Bench (Ying et al., 2024) | 15 |
| MM-Vet (Yu et al., 2023) | 15 |
| NLVR2 (Suhr et al., 2019) | 15 |
| NoCaps (Agrawal et al., 2019) | 15 |
| OCR-VQA (Mishra et al., 2019) | 15 |
| OODCV-VQA (Tu et al., 2023) | 15 |
| PathVQA (He et al., 2021) | 15 |
| PlotQA (Methani et al., 2020) | 15 |
| PMC-VQA (Zhang et al., 2023) | 15 |
| POPE (Li et al., 2023b) | 15 |
| RealWorldQA (xAI, 2024) | 15 |
| SEED-Bench (Li et al., 2023a) | 15 |
| Sketchy-VQA (Tu et al., 2023) | 15 |
| ST-VQA (Biten et al., 2019) | 15 |
| Super-CLEVR (Li et al., 2023d) | 15 |
| TabMWP (Lu et al., 2022b) | 15 |
| TextVQA (Singh et al., 2019) | 15 |
| TheoremQA (Chen et al., 2023) | 15 |
| TouchStone (Bai et al., 2023b) | 15 |
| TQA (Kembhavi et al., 2017) | 15 |
| UniGeo (Chen et al., 2022) | 15 |
| ViP-Bench (Cai et al., 2023) | 15 |
| VisDial (Das et al., 2017) | 15 |
| VisionGraph (Li et al., 2024a) | 15 |

| SOURCE DATASET | COUNT |
|---|---|
| VizWiz (Gurari et al., 2018) | 15 |
| VQA-AS (Antol et al., 2015) | 15 |
| VQA-Med (Abacha et al., 2019) | 15 |
| VQA-RAD (Lau et al., 2018) | 15 |
| VQAv2 (Goyal et al., 2017) | 15 |
| WebLINX (Lù et al., 2024) | 15 |
| Ai2D (Kembhavi et al., 2016) | 14 |
| InfographicVQA (Mathew et al., 2022) | 14 |
| MME (Fu et al., 2024) | 14 |
| OK-VQA (Marino et al., 2019) | 14 |
| SciBench (Wang et al., 2023) | 14 |
| ScienceQA (Lu et al., 2022a) | 13 |
| VisIT-Bench (Bitton et al., 2023) | 13 |
| OCRBench (Liu et al., 2024e) | 12 |
| VisualWebBench: Action Grounding (Liu et al., 2024c) | 10 |
| VisualWebBench: Action Prediction (Liu et al., 2024c) | 10 |
| VisualWebBench: Element Grounding (Liu et al., 2024c) | 9 |
| TOTAL TASKS | 66 |
| TOTAL INSTANCES | 962 |

Table 6: A full list of source datasets composing MultiSkill.