

SimpleVQA: Multimodal Factuality Evaluation for Multimodal Large Language Models

Xianfu Cheng^{1,2*}, Wei Zhang^{1*}, Shiwei Zhang^{2*}, Jian Yang^{1*†}, Xiangyuan Guan¹, Xianjie Wu¹
 Xiang Li^{1†}, Ge Zhang³, Jiaheng Liu³, Yuying Mai³, Yutao Zeng¹, Zhoufutu Wen³
 Ke Jin^{1,2}, Baorui Wang¹, Weixiao Zhou¹, Yunhong Lu³, Hangyuan Ji¹
 Tongliang Li¹, Wenhao Huang³, Zhoujun Li^{1,4}
¹Beihang University; ²Baidu Inc., China; ³M-A-P;
⁴Shenzhen Intelligent Strong Technology Co.,Ltd.
 {buaacxf, lizj}@buaa.edu.cn, zhangshiwei05@baidu.com

Abstract

The increasing application of multi-modal large language models (MLLMs) across various sectors has spotlighted the essence of their output reliability and accuracy, particularly their ability to produce content grounded in factual information (e.g. common and domain-specific knowledge). In this work, we introduce SimpleVQA, the first comprehensive multi-modal benchmark to evaluate the factuality ability of MLLMs to answer natural language short questions. SimpleVQA is characterized by 7 key features: it is based on bilingual, it covers multiple tasks and multiple scenarios, ensures high quality and challenging queries, maintains static and timeless reference answers, and is straightforward to evaluate. Our approach involves categorizing visual question-answering items into 9 different tasks around objective events or common knowledge and situating these within 9 scenario domains. Rigorous quality control processes are implemented to guarantee high-quality, concise, and clear answers, facilitating evaluation with minimal variance via an LLM-as-a-judge scoring system. Using SimpleVQA, we perform a comprehensive assessment of leading 18 MLLMs and 8 text-only LLMs, delving into their image comprehension and text generation abilities by identifying and analyzing error cases.

1. Introduction

A significant challenge in large language models (LLMs) is ensuring that LLMs [1, 25] generate factually accurate and evidence-based responses. Current state-of-the-art (SOTA)

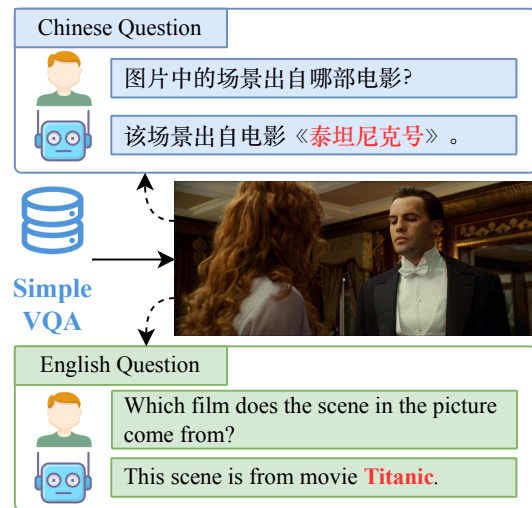


Figure 1. An example from our proposed SimpleVQA.

LLMs often produce outputs that are misleading or unsupported by evidence phenomenon known as “hallucinations” [3, 29, 43]. The generation of incorrect or unsubstantiated information remains a major barrier to the broader adoption and reliability of general-purpose AI technologies.

OpenAI proposes SimpleQA [30] to measure factuality simple and reliable with nearly 4K concise and fact-seeking questions. Further, Chinese SimpleQA [8] comprised of 3K Chinese questions spanning 6 major topics is proposed to target the Chinese language. However, the SimpleQA and Chinese SimpleQA benchmarks mainly evaluate the model capabilities in text modality, ignoring wider real-world scenarios (e.g. vision modality). For the vision modality, the research progress of the multi-modal large language models (MLLMs) is still hindered by the “hallucinations” introduced by the given images. Therefore, *The community of MLLMs has an urgent need for how to measure the simple*

* Equal Contributions.

† Corresponding Authors.

and reliable factuality introduced by the image.

To address this limitation, we develop the SimpleVQA benchmark as shown in Figure 1, where we define the factual question-answering capability of the visual language model. For the proposed factual Visual Question Answering (VQA), we collect 2,025 high-quality question-answer pairs covering 9 different scenario domains across 9 different application tasks. As a factual benchmark for a short answer, SimpleVQA has the following advantages: (1) **English and Chinese:** SimpleVQA provides general knowledge visual Q&A in both English and Chinese backgrounds, and comprehensively assesses the fact-generating capacity of MLLMs in Chinese and English communities. (2) **Multi-task division:** We divide the SimpleVQA assessment set into 9 different forms of VQA tasks according to the collected questions and different needs of pictures. (3) **Diversified scenarios:** SimpleVQA divided the VQA tasks into 9 uniquely representative scenario domains based on the category of the image and the source of the collected images. (4) **High quality:** We implement a comprehensive and rigorous quality control process to ensure the quality of questions and the accuracy of answers at SimpleVQA. (5) **Challenge:** simpleVQA focuses on factual questions that mainstream MLLMs cannot answer accurately, and cannot trace the cause of errors through the model itself. (6) **Static answers:** Following SimpleQA’s factual definition, all the standard answers provided in our benchmark don’t change over time. (7) **Easy to evaluate:** SimpleVQA’s short answers make it possible to use existing LLMs to run a judge program (called LLM-as-a-Judge [15, 34]) to quickly determine right or wrong and get an overall accuracy rate.

We systematically evaluate 18 MLLMs on SimpleVQA and create a dynamic leaderboard to show results. Further, a series of probing experiments are performed to explore the effect of the key factors for SimpleVQA. We classify the capabilities possessed by MLLMs for factual questions into two aspects, visual understanding and internalized knowledge capabilities: (1) visual understanding refers to the ability of the model to identify the subject of the question being asked in the question; and (2) internalized knowledge capabilities test whether the model has already mastered the relevant knowledge of the subject of the question being asked, and thus is able to answer the relevant question correctly after identifying that subject. Based on this definition, we added an abductive reasoning experiment to the basic assessment to help determine whether the badcase came from a lack of visual understanding ability or a lack of internalized knowledge ability by generating and labeling atomic questions (each atomic question corresponds to an atomic fact) for each VQA example.

The remarkable findings from SimpleVQA are summarized as: (1) The factual accuracy of most evaluation models in the field of visual question-answering is insufficient.

(2) The training data of MLLMs contains knowledge errors and they are overconfident in what they generate. (3) Image content understanding is still a major challenge for MLLMs to achieve improved capabilities. (4) The ability of MLLMs to internalize massive world knowledge still needs to be improved, and overcoming illusions remains a great challenge for large language models. (5) Further aligning visual core information with the knowledge stored in LLMs is also key to improving the factuality of MLLMs.

2. SimpleVQA

2.1. Overview

The SimpleVQA benchmark consists of 2,025 samples spanning 9 core tasks and 9 primary domains, with each question-image pair categorized into relevant subcategories, enabling a comprehensive evaluation of MLLMs across diverse knowledge areas. The dataset 9 tasks, including covers Logic & Science (LS), Object Identification Recognition (OIR), Time & Event (TE), Person & Emotion (PE), Location & Building (LB), Text Processing (TP), Quantity & Position Relationship (QPR), Art & Culture (AC), and Object Attributes Recognition (OAR). To ensure broad topic coverage, SimpleVQA is structured around 9 key domains: Literature, education & sports (LES), Euro-American History & Culture (EHC), Contemporary Society (CS), Engineering, Technology & Application (ETA), Film, Television & Media (FTM), Natural Science (NS), Art (AR), Chinese History & Culture (CHC), and Life (LI).

As shown in Table 1, SimpleVQA differs from existing MLLM benchmarks by focusing on factual knowledge boundaries instead of general vision-language understanding. Politically sensitive and ideological content is excluded to maintain neutrality and avoid controversy. Designed for efficiency, the dataset features concise questions and standardized answers, reducing complexity in model evaluation. All samples follow a short-answer Q&A format, enabling simple and objective assessment through direct answer matching. These refinements ensure SimpleVQA serves as a robust benchmark for evaluating MLLMs’ factual reasoning abilities.

2.2. Dataset Criteria

SimpleVQA maintains the following rigorous standards and guidelines for question, image, and answer selection, ensuring objectivity, temporal stability, and verifiability.

Question Guidelines *Clear and Unique Answers:* Questions must have a single, undisputed answer. They should precisely define scope (e.g., “Which city?” instead of “Which location?”) and specify time references (e.g., “Which year?” rather than “When?”). *Evidence-Based:* Each question must be supported by verifiable

Benchmark	Multimodal	Data Size	Language	Data Source	Domain	Factuality	Reasoning	Metric
MMBench [23]	Image&Text	2,438	Chi.&Eng.	Real World	Knowledge	×	×	MCQ Eval
CCBench [23]	Image&Text	510	Chinese	Knowledge	Knowledge	×	×	MCQ Eval
MME [13]	Image&Text	1300	English	Real World	General	×	×	TFQ Eval
MM-Vet [39]	Image&Text	200	English	Human	General	×	×	LLM-as-a-Judge
Dynamath [45]	Image&Text	5000	English	Exams	Math	×	×	Accuracy
MMM [40]	Image&Text	11.5k	English	Exams	General	×	✓	Accuracy
MMMU-Pro [41]	Image&Text	3460	English	Human&GPT	General	×	✓	Accuracy
ChineseFactEval [35]	Text Only	125	Chinese	Human	Knowledge	✓	×	LLM-as-a-Judge
AGI-Eval [44]	Text Only	8062	Chi.&Eng.	Exams	Knowledge	×	×	Accuracy
C-Eval [11]	Text Only	13,948	Chinese	Exams	Knowledge	×	×	Accuracy
SimpleQA [31]	Text Only	4,326	English	Human	Knowledge	✓	×	LLM-as-a-Judge
Chinese SimpleQA [9]	Text Only	3,000	Chinese	Human&GPT	Knowledge	✓	×	LLM-as-a-Judge
SimpleVQA (Ours)	Image&Text	2,025	Chi.&Eng.	Human&GPT	Knowledge	✓	✓	LLM-as-a-Judge

Table 1. Comparisons between our SimpleVQA and other benchmarks, where “TFQ” means True or False questions, “MCQ” means multi-choice questions, “Chi.& Eng.” means Chinese and English.

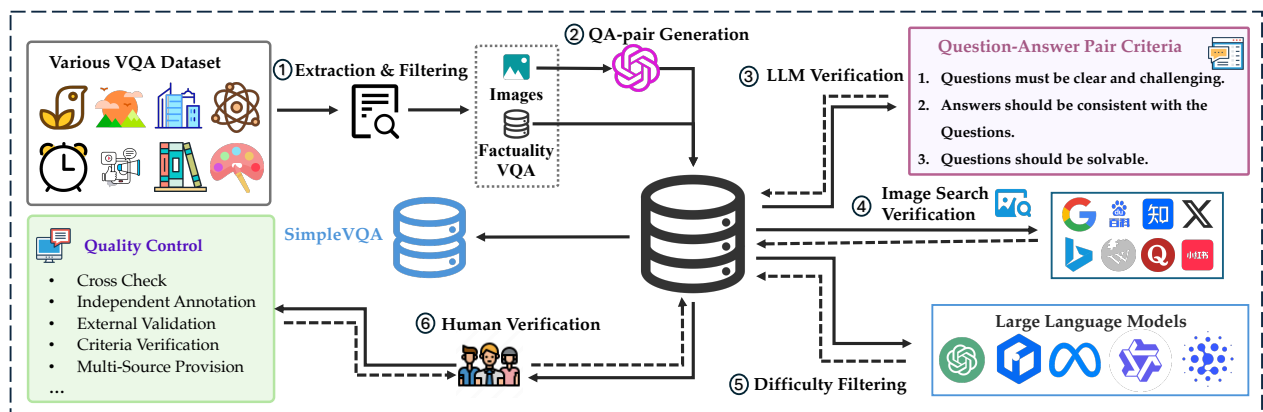


Figure 2. An overview of the data construction process of SimpleVQA.

sources. Manually annotated questions include reference links, while automatically generated ones undergo independent validation by two AI trainers. *Challenging for MLLMs*: Questions are tested on GPT-4o, GPT-4o-mini, doubao-vision-pro, and ERNIE-VL. Only those that at least one model answers incorrectly are retained; others are revised. *Answerable by August 2024*: All questions must be answerable based on knowledge available before September 1, 2024, ensuring a fair evaluation across models with similar knowledge cutoffs.

Visual Guidelines *No Direct Textual Clues*: Images must not contain text revealing the answer. *Authenticity*: Only real, unaltered images are allowed to prevent factual distortion. *Supports Question Reasoning*: Each image must provide sufficient context for answering. Manually labeled samples undergo multi-annotator verification. *Fixed Before August 2024*: Image content must be valid and confirmable before August 2024.

Answer Guidelines *Temporal Stability*: Answers must remain unchanged and unaffected by new information.

Time-sensitive topics (e.g., sports, media) should specify a timeframe rather than a general answer that may change. *Sufficiently Challenging*: Answers are tested against four high-precision MLLMs. If all models respond correctly, the question is revised to increase difficulty. *Fully Objective and Evaluable*: Answers must be precise, verifiable, and free from subjective interpretation. *Unambiguous*: Each answer must have a single, clear meaning to prevent misinterpretation.

2.3. Data Collection and Processing

Figure 2 shows an overview of the data construction process of SimpleVQA, which adheres to a meticulously designed five-step process, ensuring the highest standards of data quality and reliability.

Step 1: Seed Example Collection. SimpleVQA’s seed examples are sourced from two primary channels. First, we filter images and Q&A pairs from publicly available VQA datasets that align with factual knowledge criteria. We select MMVet (English), MME (English), Dynamath (English), MMBench_CN (Chinese), and CCBench (Chinese) due to their recent construction (post-2024) and their rele-

Statistics	Number	Statistics	Number
Data	2025	Domain Categories	9
- Chinsne(CN)	1012	- Literature, Education & Sports(LES)	13.48%
- English(EN)	1013	- Euro-American History & Culture(EHC)	12.89%
Task Categories	9	- Contemporary Society(CS)	11.51%
- Logic & Science(LS)	5.04%	- Engineering, Technology & Application(ETA)	7.95%
- Object Identification Recognition(OIR)	14.07%	- Film, Television & Media(FTM)	10.62%
- Time & Event(TE)	9.98%	- Natural Science (NS)	12.64%
- Person & Emotion(PE)	13.58%	- Art (AR)	7.65%
- Location & Building(LB)	21.53%	- Chinese History & Culture(CHC)	9.68%
- Text Processing(TP)	10.07%	- Life (LI)	13.58%
- Quantity & Position Relationship(QPR)	10.12%	Word length of Questions	
- Art & Culture(AC)	9.09%	- Maximum question length	314
- Object Attributes Recognition(OAR)	6.52%	- Minimum problem length	5

Table 2. Statistics of SimpleVQA

vance to real-world applications. Second, we collect images and relevant factual knowledge from search engines (e.g., Google, Baidu, Wikipedia), with expert annotators generating corresponding questions and answers. These data focus on entities and events across multiple domains, ensuring answers are objective, fact-based, and centered on entity recognition or attribute extraction.

Step 2: Data Enhancement and QA Pair Generation.

Once sufficient seed examples are gathered, we employ GPT-4o [12] to refine the data and generate Q&A pairs for factual categories. For multiple-choice questions (MCQs) from sources like MMbench_CN and CCBench, to ensure answer uniqueness, we use LLMs to rephrase the original question and introduce qualifiers that precisely align with the correct response. For MME, we extract the answer entity and rewrite the question based on its attributes, ensuring a one-to-one correspondence. Datasets like MMVet, Dynamath, and CCBench, which contain discrepancies from factual Q&A formats (e.g., incorrect answer options, image descriptions, or MCQ distractors), are processed using GPT-4o to align the content with factual reasoning. These refinements produce the initial version of SimpleVQA.

Step 3: LLM-Based Quality Verification. The refined dataset undergoes verification using GPT-4 to assess adherence to quality standards, ensuring answer stability, uniqueness, and question difficulty. Following LLM screening, two professional annotators conduct rigorous quality checks and refine samples as needed.

Step 4: Difficulty Screening. To maximize the dataset’s utility in model evaluation, we filter out overly simple Q&A pairs. We assess responses from four mainstream MLLMs (GPT-4o, GPT-4o-mini, Doubao-vision-pro, and ERNIE-VL). Any question correctly answered by all four models is deemed too simple and excluded from the dataset, thereby maintaining a challenging benchmark.

Step 5: Extracting Atomic Facts. To analyze visual comprehension and language alignment in MLLMs more precisely, we generate atomic questions from each SimpleVQA entry. An atomic fact represents the most funda-

mental, indivisible attribute or characteristic of an object. For instance, given the question “In what year was the person in the image born?”, the corresponding atomic question is “Who is the person in the image?”. MLLMs generate candidate answers, which are then reviewed and refined by professional annotators to ensure accuracy.

2.4. Human Annotation & Quality Control

To ensure dataset quality, we implement a rigorous manual validation process following automated data collection. Most of the collaborators participated in the necessary data annotation, and we also selected three domain experts from the collaborators. Each question is independently reviewed by two expert annotators to verify factual accuracy. If either annotator finds a question unsuitable, it is discarded. Annotators fact-check answers using authoritative sources such as Wikipedia and Baidu Encyclopedia, providing at least two supporting URLs. If their answers differ, a third expert conducts a final review to ensure consistency and correctness. Only Q&A pairs that fully align with both human evaluations and LLM-generated responses are retained.

A difficulty assessment further refines the dataset. We begin with 8,360 Q&A pairs, filtering out 22% of image-based samples that lack challenge or fail to meet predefined criteria. 1,108 pairs are removed through multi-model testing to ensure that questions pose a meaningful challenge to MLLMs. To maintain category balance, we carefully select 200 high-difficulty mathematical Q&As from 5,000 Dynamath samples, avoiding an overrepresentation of simpler factual questions. Through multiple validation rounds, we retain 2,025 high-precision Q&A pairs, accounting for 24% of the original dataset. This process ensures factual integrity, topic diversity, and appropriate difficulty levels, making SimpleVQA a robust benchmark for evaluating MLLMs’ reasoning and knowledge boundaries.

2.5. Dataset Statistics

As shown in Table 1 and 2, our SimpleVQA benchmark consists of 2,025 samples across 9 major tasks, 9 major do-

mains, and 244 image types. Examples of each category can be found in **supplementary material**. This design facilitates a comprehensive assessment of MLLMs across different domains. Regarding the distribution of domain topics and image types in SimpleVQA, 9 main domains are defined and subcategories are assigned based on each topic. Table 2 displays statistics related to SimpleVQA, including the amount of data, categorization and percentage, and the text length of the questions. In Table 1, we also compare SimpleVQA with several mainstream MLLMs’ evaluation benchmarks, which suggests that SimpleVQA is the first MLLMs’ benchmark that focuses on the evaluation of knowledge boundaries in factual categories. We excluded ideological and politically relevant data from the dataset to prevent social controversies and negative impacts. In addition, we implemented several optimizations to improve the efficiency of the evaluation. The dataset features concise questions and standardized answers, minimizing the input and output markers required for GPT assessment. In addition, all examples are in short-answer question-and-answer (QA) format, and they can be assessed by simple matching.

3. Experiments

3.1. Experimental Setup

We implement a standardized protocol across all experiments, maintaining a consistent prompt format throughout. All temperature and sampling parameters are configured according to each LLM’s official recommendations or default settings. For evaluation and benchmark construction, we employ GPT-4o as the primary model due to its robust performance in multimodal tasks.

3.2. Baseline Models

We evaluate 18 representative MLLMs in total, comprising 8 closed-source and 10 open-source models, which provide a diverse evaluation of model capabilities across different architectures and training paradigms. The closed-source models include GPT-4o and 4o-mini, Doubao-pro-128k and 32k, Gemini-2.0-flash, Claude-3.5-Sonnet, Qwen-Max, ERNIE-VL. The open-source models cover a wide range of frameworks, including InternVL2.5 and In-

ternVL2, Qwen2.5-VL and Qwen2-VL, Janus-pro-7B.

3.3. Evaluation Metrics

The SimpleVQA benchmark employs a set of evaluation metrics to assess the accuracy, reliability, and consistency of the model’s predictions. These metrics include: (1) **Correct (CO)** evaluates whether the predicted answer matches the reference answer exactly without any contradiction. (2) **Not Attempted (NA)** identifies cases where the model does not attempt to answer, ensuring no contradictions are present. (3) **Incorrect (IN)** flags instances where the predicted answer contradicts the reference answer, even if resolved. (4) **Correct Given Attempted (CGA)** measures the proportion of correct answers among those attempted by the model, reflecting its performance when engaged. (5) **F-score** computes the harmonic mean between “Correct” and “Correct Given Attempted”, providing a balanced evaluation that combines accuracy and attempt success.

3.4. Main Results

Results on Overall and Different Tasks Table 3 presents the performance of MLLMs on SimpleVQA, highlighting their F-scores across different tasks. Among the closed-source models, Gemini-2.0-Flash demonstrates superior performance, particularly in tasks such as LS, OIR, TE, PE, LB, TP, and QPR, achieving state-of-the-art (SOTA) scores across these task categories. Claude-3.5-Sonnet, GPT-4o, and ERNIE also exhibit strong performance. Open-source models, including InternVL2.5-78B-MPO and Qwen2.5-VL-72B-Instruct, show competitive results, though slightly below the leading closed-source models. Notably, Qwen2.5-VL-72B-Instruct outperforms most mainstream closed-source MLLMs in several aspects.

Results on Cross-lingual and Different Domains Table 4 lists the performance of various closed-source and open-source MLLMs on SimpleVQA, highlighting their Q&A performance across languages and across different scenarios. Among the closed-source models, Gemini-2.0-Flash and ERNIE-VL show strong Chinese quizzing ability; Gemini-2.0-Flash and Claude-3.5-Sonnet show strong English quizzing ability; especially on the scale evaluation with scenarios as the categorization, the three models mentioned above achieve almost all the SOTA. open-source models InternVL2.5-78B-MPO and Qwen2.5-VL-

<https://platform.openai.com/docs/models#gpt-4o>; <https://platform.openai.com/docs/models#gpt-4o-mini>
<https://www.volcengine.com/product/ark>
<https://deepmind.google/technologies/gemini/flash/>
<https://www.anthropic.com/news/claude-3-5-sonnet>
<https://huggingface.co/spaces/Qwen/Qwen-Max>
<https://yiyao.baidu.com/>

https://huggingface.co/OpenGVLab/InternVL2_5-78B-MPO; <https://huggingface.co/OpenGVLab/InternVL2-Llama3-76B>
<https://huggingface.co/Qwen/Qwen2.5-VL-72B-Instruct>; <https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct>; <https://huggingface.co/Qwen/Qwen2-VL-72B-Instruct>
<https://huggingface.co/deepseek-ai/Janus-Pro-7B>

Models	Overall results					F-score on 9 task categories								
	CO	NA	IN	CGA	F-score	LS	OIR	TE	PE	LB	TP	QPR	AC	OAR
Closed-source Multi-modal Large Language Models														
GPT-4o	47.2	7.8	45.0	51.2	49.1	58.7	53.4	48.1	36.4	44.3	57.5	61.8	35.1	61.0
GPT-4o-mini	35.5	10.0	54.5	39.4	37.3	30.0	44.4	38.5	29.6	36.2	42.1	45.9	25.8	40.6
Doubao-vision-pro-128k	39.7	20.8	39.5	50.1	44.3	48.3	53.5	53.4	39.6	35.0	38.5	46.9	35.9	62.9
Doubao-vision-pro-32k	25.4	23.6	51.4	32.7	28.3	24.2	37.9	18.0	31.6	21.3	39.2	32.9	18.6	36.6
Gemini-2.0-flash	52.8	6.0	41.2	56.1	54.4	63.7	60.9	54.3	55.0	44.3	61.5	65.7	33.9	64.4
Claude-3.5-Sonnet	48.5	9.8	41.8	53.7	50.9	57.1	53.9	52.1	29.7	47.2	58.4	62.7	46.7	60.1
Qwen-Max	25.4	15.3	59.3	30.0	27.5	15.1	33.7	13.5	30.1	27.7	36.0	34.4	12.1	36.4
ERNIE-VL	46.5	9.5	44.0	51.4	48.8	49.0	55.9	48.3	40.7	40.7	54.4	59.7	33.6	70.8
Open-source Multi-modal Large Language Models														
InternVL2.5-78B-MPO	45.4	5.9	48.6	48.3	46.8	57.4	54.6	50.3	26.4	34.4	57.0	65.8	32.8	72.3
InternVL2.5-78B	41.5	7.7	50.8	45.0	43.2	49.5	49.4	45.0	28.4	31.0	49.6	63.7	32.0	65.2
InternVL2-Llama3-76B	35.7	8.4	55.9	38.9	37.2	34.7	43.5	35.6	26.2	28.4	44.3	53.5	29.0	54.1
InternVL2.5-38B-MPO	42.9	5.4	51.8	45.3	44.0	51.7	52.5	45.7	22.1	34.0	52.0	67.8	32.8	60.4
InternVL2.5-26B-MPO	39.9	7.6	52.5	43.2	41.5	43.4	44.5	47.7	27.1	31.7	45.9	58.6	29.0	68.0
InternVL2.5-8B-MPO	33.6	7.5	58.9	36.3	34.9	45.1	37.4	30.9	19.1	26.0	44.3	52.2	26.4	56.2
Qwen2.5-VL-72B	49.4	5.4	45.2	52.2	50.8	57.7	50.8	51.0	38.6	51.6	57.8	65.8	29.8	62.8
Qwen2-VL-72B-Instruct	44.7	10.3	45.0	49.8	47.1	48.8	51.9	46.3	37.9	38.5	55.3	63.0	35.9	59.7
Qwen2.5-VL-7B-Instruct	43.2	5.1	51.7	45.6	44.3	38.6	52.7	37.6	32.7	41.0	52.6	53.6	39.2	56.7
Janus-pro-7B	31.3	10.6	58.1	35.0	33.0	27.0	43.4	26.5	23.7	28.2	24.0	50.4	36.2	42.3

Table 3. Results of different models on SimpleVQA. For metrics, CO, NA, IN, and CGA denote “Correct”, “Not attempted”, “Incorrect”, and “Correct given attempted”, respectively. We report the scores across different tasks, including “Logic & Science (LS)”, “Object Identification Recognition (OIR)”, “Time & Event (TE)”, “Person & Emotion (PE)”, “Location & Building (LB)”, “Text Processing (TP)”, “Quantity & Position Relationship (QPR)”, “Art & Culture (AC)”, and “Object Attributes Recognition (OAR)”.

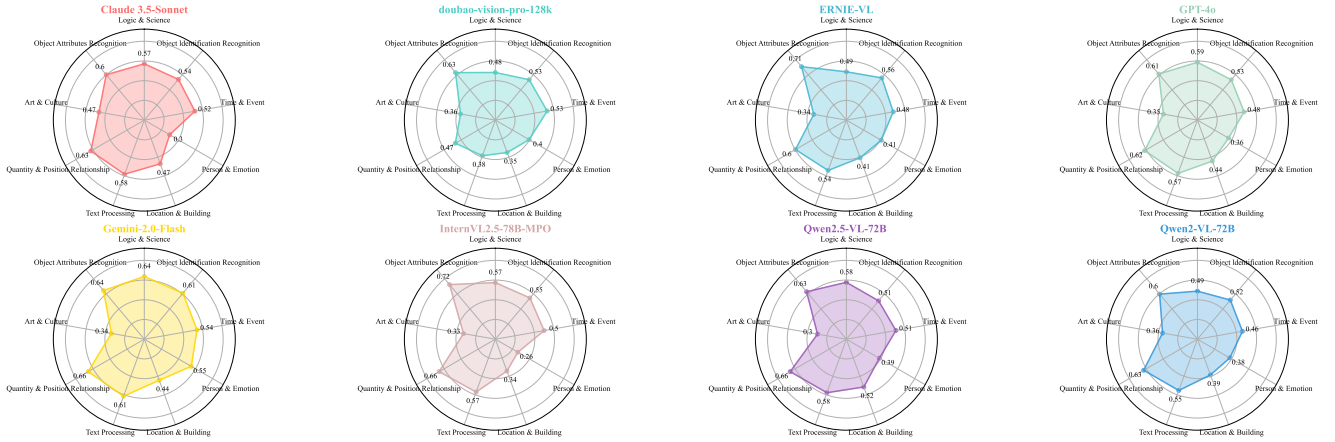


Figure 3. F-score results for eight different models across nine task categories.

72B-Instruct, respectively, achieved the highest performance, albeit slightly lower than the top closed-source models. It is worthwhile to note that Qwen2.5-VL-72B-Instruct achieved the highest score for the open source model on the English review and was reviewed for strong performance on several domains.

Confidence Analysis The study of confidence analysis evaluates the consistency between an MLLM’s expressed

confidence and the accuracy of its answers by requiring models to report confidence levels with their responses. As illustrated in Figure 4, two key patterns emerge: first, models demonstrate higher confidence when more reference materials are available, irrespective of architectural differences; second, while higher confidence generally correlates with greater accuracy, we observe numerous instances of high-confidence errors. This finding suggests that MLLMs exhibit overconfidence in their knowledge representation

Models	Chinese partial results			English partial results			F-score on 9 domains categories								
	CO	CGA	F-score	CO	CGA	F-score	LES	EHC	CS	ETA	FTM	NS	AR	CHC	LI
Closed-source Multi-modal Large Language Models															
GPT-4o	48.7	51.6	50.1	45.7	50.7	48.1	47.0	37.5	58.2	62.0	50.5	61.7	30.3	29.6	58.8
GPT-4o-mini	33.0	36.0	34.4	38.2	43.1	40.5	36.0	25.4	44.4	41.5	41.1	40.6	22.9	26.6	52.3
Doubao-vision-pro-128k	51.1	56.4	53.6	28.5	41.8	33.9	51.8	23.7	50.1	52.2	49.0	44.3	32.6	44.9	48.7
Doubao-vision-pro-32k	29.0	37.8	32.8	15.6	26.2	19.6	34.9	13.4	33.0	38.2	29.3	23.4	15.2	29.1	37.4
Gemini-2.0-flash	54.8	57.9	56.3	50.7	54.3	52.5	54.1	35.4	59.9	67.5	70.9	64.6	29.2	36.3	64.6
Claude-3.5-Sonnet	48.0	50.6	49.3	49.2	57.0	52.8	49.1	42.7	52.7	56.5	48.1	61.1	38.7	34.4	66.5
Qwen-Max	25.2	29.4	27.1	25.6	30.6	27.9	28.1	21.0	35.4	34.8	31.6	20.0	14.2	11.9	44.1
ERNIE-VL	54.0	55.3	54.6	40.7	44.9	42.7	55.9	32.9	60.0	54.6	42.8	50.7	27.8	45.4	59.4
Open-source Multi-modal Large Language Models															
InternVL2.5-78B-MPO	51.7	54.7	53.1	39.2	41.8	40.5	54.2	22.5	60.2	53.6	35.5	56.7	27.5	37.9	63.6
InternVL2.5-78B	46.7	49.2	47.9	36.3	40.6	38.3	48.3	18.6	55.2	56.0	33.0	53.9	25.6	33.7	57.5
InternVL2-Llama3-76B	34.2	37.4	35.7	37.1	40.4	38.7	38.3	19.0	44.0	44.4	35.4	45.6	20.4	21.4	57.1
InternVL2.5-38B-MPO	48.0	49.5	48.8	37.7	40.9	39.2	52.3	21.0	54.2	59.5	29.3	51.5	27.9	31.6	61.6
InternVL2.5-26B-MPO	45.5	47.3	46.3	34.4	38.8	36.4	46.9	20.6	54.5	47.6	32.4	51.6	20.5	34.4	54.3
InternVL2.5-8B-MPO	35.8	38.5	37.1	31.4	34.0	32.7	36.0	14.0	43.7	43.2	26.4	46.4	20.3	22.5	53.5
Qwen2.5-VL-72B-Instruct	48.0	50.4	49.2	50.7	54.0	52.3	49.4	45.2	64.3	62.4	53.6	58.3	20.9	30.3	61.8
Qwen2-VL-72B-Instruct	46.1	48.6	47.3	43.3	51.2	46.9	45.2	30.4	54.0	57.2	51.6	53.5	29.4	29.7	65.2
Qwen2.5-VL-7B-Instruct	41.9	44.0	42.9	44.5	47.1	45.8	42.9	42.8	54.9	49.8	40.4	46.4	25.9	30.3	56.9
Janus-pro-7B	29.5	32.1	30.7	33.2	38.1	35.5	30.5	21.5	40.2	40.8	26.2	37.1	26.8	15.3	53.2

Table 4. Results of different models on SimpleVQA. For metrics, CO, NA, IN, and CGA denote “Correct”, “Not attempted”, “Incorrect”, and “Correct given attempted”, respectively. SimpleVQA is structured around nine key domains: “Literature, education & sports (LES)”, “Euro-American History & Culture (EHC)”, “Contemporary Society (CS)”, “Engineering, Technology & Application (ETA)”, “Film”, “Television & Media (FTM)”, “Natural Science (NS)”, “Art (AR)”, “Chinese History & Culture (CHC)”, and “Life (LI)”.

and comprehension capabilities.

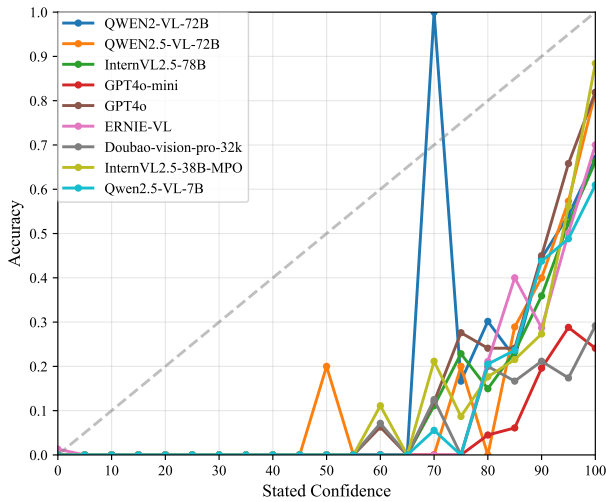


Figure 4. Calibration of LLMs based on their stated confidence. The x-axis represents the confidence level of the LLMs, and the y-axis represents the accuracy.

Results on Different LLMs To validate the benchmark’s reliance on visual information, we conduct experiments with eight mainstream LLMs using text-only inputs (without images). These models include well-known LLMs

Models	CO	IN	NA	CGA
o1-preview	3.65%	22.12%	74.22%	14.16%
o1-mini	4.1%	19.51%	76.4%	17.36%
DeepSeek-R1	10.08%	53.6%	36.32%	15.82%
Qwen-Max	7.6%	70.77%	21.63%	9.69%
GPT-4o	5.23%	31.6%	63.16%	14.2%
GPT-4o-mini	6.47%	47.65%	45.88%	11.95%
Claude-3.5-Sonnet2	2.02%	7.21%	90.77%	21.88%
Gemini-2.0-flash	7.06%	61.33%	31.6%	10.32%

Table 5. The CO, NA, IN, and CGA results for 8 LLMs across simpleVQA without image input.

such as o1-Preview, o1-mini, and DeepSeek-R1. Table 5 presents the experimental results. We specifically design visual questions that theoretically require image inputs to be answered correctly. Despite the absence of visual information, LLMs achieve minimal accuracy on these tasks, with DeepSeek-R1 demonstrating notably higher performance in inferring correct answers from textual context alone.

4. Further Analysis

Based on SimpleVQA, we conduct a further evaluation of MLLMs, revealing significant factual accuracy issues. We

<https://platform.openai.com/docs/models#o1>
<https://platform.openai.com/docs/models#o1-mini>
<https://huggingface.co/deepseek-ai/DeepSeek-R1>

also perform a deep analysis of these factual problems by examining both the visual understanding capabilities and text generation mechanisms of MLLMs, which provides insights for future model optimization.

First, we identify the three most robust MLLMs through systematic evaluation. For each VQA task where an MLLM provides an incorrect response, we use a carefully designed prompt to decompose the question into an atomic problem focused on basic content recognition. Each atomic problem corresponds to a fundamental atomic fact that, when identified, converts the original question into a purely factual text-based query. When an MLLM fails to answer even these simplified atomic queries correctly, we attribute this failure to insufficient visual understanding of the image. Next, we distinguish between original questions that are already atomic in nature and those that are not. We systematically identify cases where the derived atomic questions differ from the original questions and compile them into a test set called the Complex Fact Question (CFQ) set. This dataset is specifically designed to evaluate whether model performance improves when atomic facts are explicitly provided. In a follow-up experiment, we augment each CFQ query by incorporating the answer to its corresponding atomic question as an explicit hint. We then reassess model performance with these augmented queries. When models still provide incorrect answers despite having access to the relevant atomic facts, we attribute such failures to deficiencies in background knowledge rather than visual understanding.

Table 6 presents the results of the CFQ experiments. From our complete dataset, we select 570 challenging CFQ examples to evaluate both visual understanding and knowledge internalization capabilities. In addition, another 90 questions are selected from which none MLLMs can answer correctly, to test the performance of mainstream LLMs in deep thinking. We test pure language models (LLMs) including o1-preview, o1-mini, and DeepSeek-R1, as well as multimodal models (MLLMs) such as GPT-4o, Qwen2.5-VL-72B-Instruct, and InternVL2.5-78B-MPO. Our findings reveal several key insights. First, LLMs, despite their reasoning capabilities, fail to effectively utilize atomic facts without accompanying images. Second, MLLMs demonstrate a significant disparity between their visual understanding abilities and knowledge internalization capabilities. While MLLMs show marginally better knowledge retention in relation to visual content, substantial room for improvement remains. Finally, the performance patterns observed when MLLMs answer atomic questions suggest potential for enhancing visual understanding through targeted training approaches.

5. Related Works

Multimodal Benchmarks. Recent vision-language benchmarks have been developed to assess models’ capabilities

Models	Origin	Atomic	Atomic-Given	Fix-Badcase
o1-preview	-	-	62.74%	13.33% (90)
o1-mini	-	-	51.49%	10.00% (90)
DeepSeek-R1	-	-	55.01%	17.78% (90)
Qwen-Max	-	-	54.83%	5.56% (90)
GPT-4o	56.24%	56.94%	61.69%	35.60% (250)
Qwen2.5-VL-72B-Instruct	51.67%	59.58%	64.15%	36.25% (276)
InternVL2.5-78B-MPO	55.36%	55.71%	69.95%	43.53% (255)

Table 6. CFQ experiment in SimpleVQA. “Origin”: original Q&A accuracy; “Atomic”: atomic Q&A accuracy; “Atomic-Given”: original Q&A with atomic facts; “Fix-Badcase”: corrected responses with atomic facts. Sample sizes are in parentheses.

in integrating visual and textual information across various tasks [14, 32, 33, 42], including OCR [4, 6], spatial awareness [17], multimodal representation learning and retrieval [5, 19, 20], and reasoning skills [24, 26]. For example, MMBench [22] employs multiple-choice tasks in both Chinese and English, covering a wide range of domains. MMMU [40] focuses on complex vision-language tasks, particularly those requiring advanced multimodal reasoning. MMStar [2] utilizes multi-task evaluations to test models’ ability to fuse different modalities.

Factuality Benchmarks. Factuality refers a model’s capacity to produce content consistent with factual knowledge, encompassing commonsense, general world knowledge, and domain-specific information. The capability can be evaluated through comparison against authoritative references, including Wikipedia and academic literature. Recently, Various benchmarks have been developed to evaluate factuality in LLMs [11, 16, 21, 27, 28, 36–38, 44]. For example, MMLU [10] assesses multitask accuracy across 57 diverse tasks. HaluEval [18] explores the propensity of LLMs to produce hallucinations or false information. SimpleQA [31] and Chinese SimpleQA [7] have been proposed to measure the short-form factuality in LLMs.

6. Conclusion

This paper presents SimpleVQA, the first bilingual visual question-answering benchmark for evaluating fact-based reasoning in MLLMs. SimpleVQA features seven key attributes: Chinese-English bilingualism, multi-task/multi-scene adaptability, high-quality content, challenging queries, static design, and evaluation efficiency. Our comprehensive evaluation of 18 MLLMs and 8 LLMs demonstrates its effectiveness in assessing fact-based reasoning capabilities. We also introduce the calibration methodology to evaluate MLLMs’ visual comprehension and visual-linguistic alignment through atomic question decomposition of factual queries. SimpleVQA establishes a robust benchmark for assessing factuality and advancing the development of reliable MLLMs.

7. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62276017, 62406033, U1636211, 61672081), and the State Key Laboratory of Complex & Critical Software Environment (Grant No. SKLCCSE-2024ZX-18).

References

- [1] AI@Meta. Llama 3 model card. 2024. 1
- [2] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 8
- [3] Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, et al. Evaluating hallucinations in chinese large language models. *arXiv preprint arXiv:2310.03368*, 2023. 1
- [4] XianFu Cheng, YanQing Yao, Liying Zhang, Ao Liu, and Zhoujun Li. An improved stochastic gradient descent algorithm based on rényi differential privacy. *International Journal of Intelligent Systems*, 37(12):10694–10714, 2022. 8
- [5] Xianfu Cheng, Hang Zhang, Jian Yang, Xiang Li, Weixiao Zhou, Kui Wu, Fei Liu, Wei Zhang, Tao Sun, Tongliang Li, et al. Xformparser: A simple and effective multimodal multilingual semi-structured form parser. *arXiv preprint arXiv:2405.17336*, 2024. 8
- [6] Xianfu Cheng, Weixiao Zhou, Xiang Li, Jian Yang, Hang Zhang, Tao Sun, Wei Zhang, Yuying Mai, Tongliang Li, Xiaoming Chen, et al. Svipt: Fast and efficient scene text recognition with vision permutable extractor. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 365–373, 2024. 8
- [7] Yancheng He, Shilong Li, Jiaheng Liu, Yingshui Tan, Weixun Wang, Hui Huang, Xingyuan Bu, Hangyu Guo, Chengwei Hu, Boren Zheng, Zhuoran Lin, Xuepeng Liu, Dekai Sun, Shirong Lin, Zhicheng Zheng, Xiaoyong Zhu, Wenbo Su, and Bo Zheng. Chinese simpleqa: A chinese factuality evaluation for large language models, 2024. 8
- [8] Yancheng He, Shilong Li, Jiaheng Liu, Yingshui Tan, Weixun Wang, Hui Huang, Xingyuan Bu, Hangyu Guo, Chengwei Hu, Boren Zheng, et al. Chinese simpleqa: A chinese factuality evaluation for large language models. *arXiv preprint arXiv:2411.07140*, 2024. 1
- [9] Yancheng He, Shilong Li, Jiaheng Liu, Yingshui Tan, Weixun Wang, Hui Huang, Xingyuan Bu, Hangyu Guo, Chengwei Hu, Boren Zheng, et al. Chinese simpleqa: A chinese factuality evaluation for large language models. *arXiv preprint arXiv:2411.07140*, 2024. 3
- [10] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 8
- [11] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*, 2023. 3, 8
- [12] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 4
- [13] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024. 3
- [14] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. *arXiv preprint arXiv:2410.14669*, 2024. 8
- [15] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv: 2411.16594*, 2024. 2
- [16] Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmu: Measuring massive multitask language understanding in chinese, 2023. 8
- [17] Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. *arXiv preprint arXiv:2501.08282*, 2025. 8
- [18] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models, 2023. 8
- [19] Xiang Li, Xianfu Cheng, Dezhuang Miao, Xiaoming Zhang, and Zhoujun Li. Tf-mamba: Text-enhanced fusion mamba with missing modalities for robust multimodal sentiment analysis. *arXiv preprint arXiv:2505.14329*, 2025. 8
- [20] Xiang Li, Haijun Zhang, Zhiqiang Dong, Xianfu Cheng, Yun Liu, and Xiaoming Zhang. Learning fine-grained representation with token-level alignment for multimodal sentiment analysis. *Expert Systems with Applications*, 269:126274, 2025. 8
- [21] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, 2022. Association for Computational Linguistics. 8
- [22] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 8

- [23] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 3
- [24] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 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, 2019. 8
- [25] OpenAI. Gpt-4 technical report. *PREPRINT*, 2023. 1
- [26] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022. 8
- [27] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. 8
- [28] Yingshui Tan, Boren Zheng, Baihui Zheng, Kerui Cao, Huiyun Jing, Jincheng Wei, Jiaheng Liu, Yancheng He, Wenbo Su, Xiangyong Zhu, and Bo Zheng. Chinese safetyqa: A safety short-form factuality benchmark for large language models, 2024. 8
- [29] SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024. 1
- [30] Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. 2024. [URL https://api.semanticscholar.org/CorpusID/273877483](https://api.semanticscholar.org/CorpusID/273877483). 1
- [31] Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. 2024. 3, 8
- [32] Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhui Chen, et al. Scimmir: Benchmarking scientific multi-modal information retrieval. *arXiv preprint arXiv:2401.13478*, 2024. 8
- [33] Siwei Wu, Kang Zhu, Yu Bai, Yiming Liang, Yizhi Li, Haoning Wu, Jiaheng Liu, Ruiibo Liu, Xingwei Qu, Xuxin Cheng, et al. Mmra: A benchmark for multi-granularity multi-image relational association. *arXiv preprint arXiv:2407.17379*, 2024. 8
- [34] Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xeron Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, et al. Tablebench: A comprehensive and complex benchmark for table question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 25497–25506, 2025. 2
- [35] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023. 3
- [36] Jian Yang, Jiaxi Yang, Ke Jin, Yibo Miao, Lei Zhang, Liquan Yang, Zeyu Cui, Yichang Zhang, Binyuan Hui, and Junyang Lin. Evaluating and aligning codellms on human preference. *arXiv preprint arXiv:2412.05210*, 2024. 8
- [37] Jian Yang, Jiajun Zhang, Jiaxi Yang, Ke Jin, Lei Zhang, Qiyao Peng, Ken Deng, Yibo Miao, Tianyu Liu, Zeyu Cui, et al. Execrepobench: Multi-level executable code completion evaluation. *arXiv preprint arXiv:2412.11990*, 2024.
- [38] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering, 2018. 8
- [39] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 3
- [40] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 3, 8
- [41] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. 3
- [42] 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. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *ArXiv*, abs/2401.11944, 2024. 8
- [43] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren’s song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023. 1
- [44] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models, 2023. 3, 8
- [45] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*, 2024. 3