CMMU: A Benchmark for Chinese Multi-modal Multi-type Question Understanding and Reasoning

Zheqi He[1 *], **Xinya Wu**[1 *], **Pengfei Zhou**[1,2 †], **Richeng Xuan**[1], **Guang Liu**[1 ‡], **Xi Yang** [1 ‡], **Qiannan Zhu** [3], **Hua Huang** [3]

[1]Beijing Academy of Artificial Intelligence
[2]Beijing University of Post and Telecommunication
[3]Beijing Normal University
{zqhe, yxwu, rcxuan,liuguang,yangxi}@baai.ac.cn, zhoupengfei@bupt.edu.cn, {zhuqiannan,huahuang}@bnu.edu.cn

Abstract

Multi-modal large language models(MLLMs) have achieved remarkable progress and demonstrated powerful knowledge comprehension and reasoning abilities. However, the mastery of domain-specific knowledge, which is essential for evaluating the intelligence of MLLMs, continues to be a challenge. Current multi-modal benchmarks for domain-specific knowledge concentrate on multiple-choice questions and are predominantly available in English, which imposes limitations on the comprehensiveness of the evaluation. To this end, we introduce **CMMU**, a novel benchmark for multi-modal and multi-type question understanding and reasoning in Chinese. CMMU consists of 3,603 questions in 7 subjects, covering knowledge from primary to high school. The questions can be categorized into 3 types: multiple-choice, multiple-response, and fill-in-the-blank, bringing greater challenges to MLLMs. In addition, we propose a rigorous evaluation strategy called **ShiftCheck** for assessing multiple-choice questions. The strategy aims to reduce position bias, minimize the influence of randomness on correctness, and perform a quantitative analysis of position bias. We evaluate seven open-source MLLMs along with GPT4-V, Gemini-Pro, and Qwen-VL-Plus. The results demonstrate that CMMU poses a significant challenge to the recent MLLMs.

1 Introduction

Currently, multi-modal large language models (MLLM) like GPT-4[OpenAI, 2023], Gemini[Team *et al.*, 2023], LLaVA[Liu *et al.*, 2023a] and Qwen-VL [Bai *et al.*, 2023] have showed powerful abilities in this field of multi-model. At the same time, the ability to evaluate MLLMs more rationally and comprehensively is receiving increasing attention. Researchers have made many efforts to address this problem. Datasets like MMBench [Fu *et al.*, 2023], MME [Fu *et al.*, 2023], and SEED-Bench [Li *et al.*, 2023b; Li *et al.*, 2023a] evaluate models through a diverse range of questions, ranging from perception to reasoning abilities. However, these datasets primarily access common-scene knowledge more than domain-specific knowledge. The recently introduced GAIA benchmark [Mialon *et al.*, 2023] focuses on testing fundamental abilities like reasoning, multi-modal processing, and general tool use. However, GAIA also presents certain limitations. It primarily tests tasks that are conceptually simple for humans, which may not fully capture the complex problem-solving capabilities required in some specialized domains.

In addition to the above benchmarks, alternative evaluation datasets containing questions from textbooks and other educational materials are proposed to evaluate domain-specific knowledge, which are inspired by human exams for measuring knowledge levels and selecting talents. For instance, ScienceQA[Lu *et al.*, 2022] is a dataset that evaluates the scientific knowledge of models, while MMMU[Yue *et al.*, 2023] assesses university-level knowledge. These two datasets only contain English questions, while some datasets, such as M3Exam [Zhang *et al.*, 2023], turn attention to the multilingual setting. However, the above benchmarks mainly focus on multiple-choice questions, which limits the comprehensiveness of evaluation. Multiple-choice questions cannot evaluate the text generation abilities of the models, as the models only need to choose the correct answer from a few existing options. Meanwhile, the models may obtain correct answers through guessing, which could impact the accuracy of the evaluation. Therefore, there is a need for a diversified and comprehensive benchmark to evaluate the understanding and reasoning abilities of MLLMs.

To bridge the dataset gap, we introduce a novel benchmark, CMMU, for multi-modal and muli-type question understanding and reasoning in Chinese. CMMU encompasses multi-modal content across 7 subjects. Every question requires the model to combine image and text content to generate a comprehensive response. While CMMU shares similarities with datasets like ScienceQA and M3Exam [Zhang

*Equal contribution.

†This work was done when Zhou was an intern at Beijing Academy of Artificial Intelligence.

‡Corresponding authors.

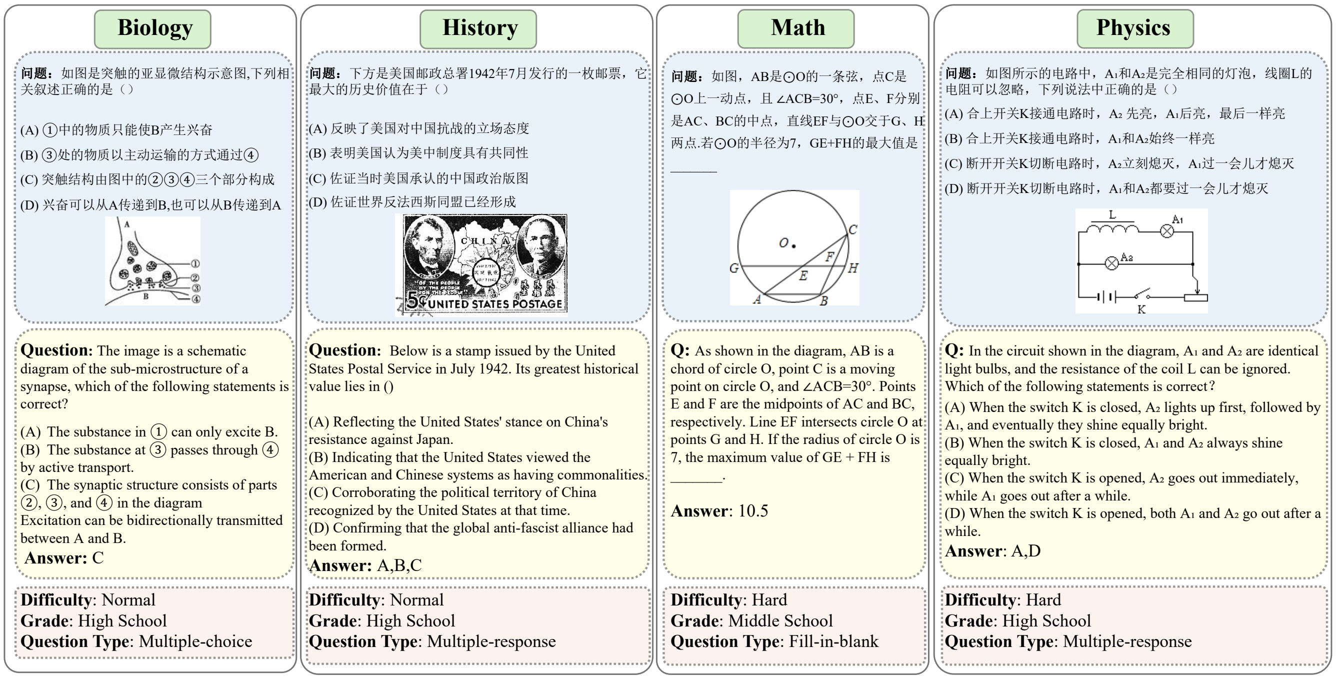


Figure 1: Some examples in MMCU. We provide Chinese examples and their corresponding English translations.

et al., 2023], it offers a broader range of question types. Previous datasets only have multiple-choice questions, while CMMU offers a wider variety of question types, including multiple-choice, multiple-response, and fill-in-the-blank questions, which poses a more significant challenge to the comprehension abilities of MLLM. In addition, to mitigate the position bias[Zheng *et al.*, 2023] in LLM and ensure genuine correctness rather than guessing, inspired by CircularEval [Liu *et al.*, 2023c], we adopt a **ShiftCheck** approach for multiple-choice questions. Specifically, we cycle through the position of options to ensure that the answer can appear at any position with equal probability, aim at reducing position bias, minimizing the influence of randomness on correctness, and performing a quantitative analysis of position bias. We evaluate 11 models using the CMMU benchmark, and the results indicate that CMMU presents a significant challenge to current MLLMs.

To sum up, our contributions are as follows:

- We present a novel benchmark of multi-modal and multi-type questions in Chinese, featuring a wider variety of question types, including multiple-choice, multiple-response, and fill-in-the-blank questions.
- We evaluate 10 models and analyze their performances in Chinese language proficiency and multi-modal comprehension.
- We propose the ShiftCheck strategy, which is designed to rigorously evaluate multiple-choice questions and conduct a quantitative analysis of position bias in MLLMs.

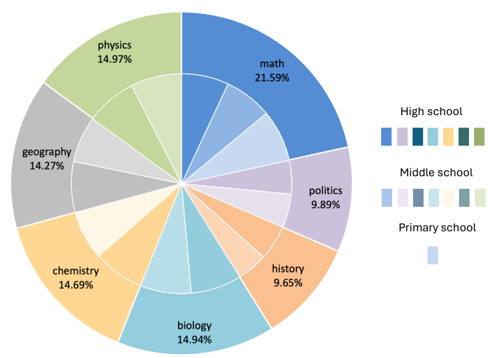


Figure 2: Distribution of the number of questions in each subject and grades.

2 Related Work

2.1 Multi-modal Benchmarks

There are already many datasets for evaluating the multi-modal capabilities of models. For instance, datasets like VQAv2 [Goyal *et al.*, 2017], TDIUC [Kafle and Kanan, 2017], TextVQA [Singh *et al.*, 2019], and GQA [Hudson and Manning, 2019] are used in visual question answering tasks, while COCO [Lin *et al.*, 2014], NoCaps [Agrawal *et al.*, 2019], and Flickr30K [Plummer *et al.*, 2015] are employed in image captioning tasks. Additionally, Visual7w [Zhu *et al.*, 2016] and RefCOCO [Kazemzadeh *et al.*, 2014] are commonly utilized for visual grounding purposes. With the rapid development of multi-modal large language models, researchers have achieved good results on these datasets. We require more extensive data to evaluate MLLMs, and there

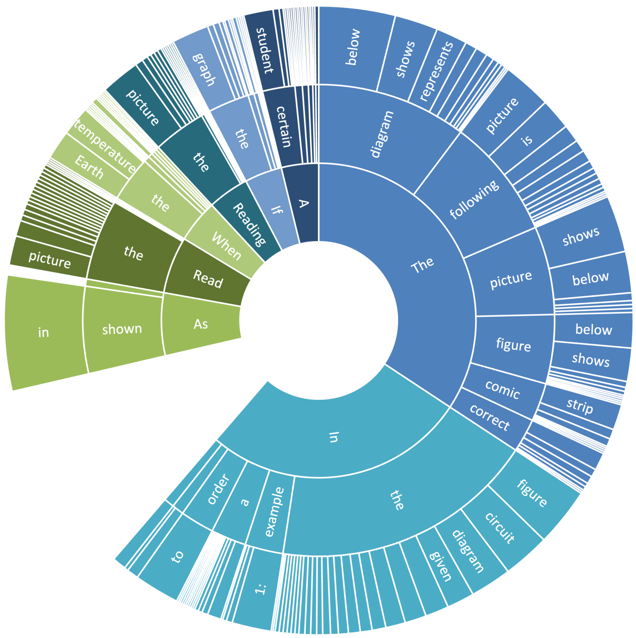


Figure 3: The first three words of questions in CMMU. We have translated it into English.

Static	**Number**
Total Questions	3,603
Validation:Test	1,800:1,803
Subjects	7
Questions with a explanation	2,585 (71.75%)
Difficulties: Normal	2,885 (80.07%)
Difficulties: Hard	718 (19.93%)
Multiple-choice question	2,710 (75.22%)
Multiple-response question	254 (7.05%)
Fill-in-the-blank question	639 (17.74%)
*Sub-questions	1,632
Primary school	250 (6.90%)
Middle school	1,697 (47.19%)
High school	1,656 (45.96%)
Average question length	72.15
Average sub-question length	43.91
Average choice length	14.47
Average explanation length	311.67

Table 1: Detailed statistics of the CMMU

have been recent studies evaluating models from various perspectives. LVLM-eHub [Xu *et al.*, 2023] collects 47 existing benchmarks and evaluates 6 types of capabilities of MLLMs, however, it does not create any new benchmarks. MME [Fu *et al.*, 2023] comprehensively measures the perception and cognition abilities of models. However, its question types are simplistic, merely requiring *yes* or *no* responses. MMBench [Liu *et al.*, 2023c] and SEED-Bench [Li *et al.*, 2023b; Li *et al.*, 2023a] contain many multiple-choice questions covering various ability dimensions, but these datasets mainly consist of common-sense questions and do not require lots of domain-specific knowledge and complex reasoning. To enhance the evaluation of domain-specific knowledge, ScienceQA [Lu *et al.*, 2022] was introduced. This dataset encompasses a wide range of science topics from elementary and high school curricula. MMMU [Yue *et al.*, 2023] is designed to evaluate college-level subject knowledge, questions of CMMU are collected from college exams and textbooks, and many of them require expert-level skills. M3Exam [Zhang *et al.*, 2023] is a multilingual and multi-modal benchmark designed to evaluate domain knowledge and problem-solving skills, and it spans seven languages. However, less than one-third of the questions include images.

2.2 Multi-modal Large Language Models

Benefiting from the development of large language models(LLM) like GPT [Macfarlane, 2023], LLaMA [Touvron *et al.*, 2023] and Vicuna [Chiang *et al.*, 2023], MLLMs have made significant improvements. Many works integrate additional modal inputs on LLM and bridge the modality gap between vision and language, and the majority of MLLMs possess multilingual capabilities. BLIP-2 [Li *et al.*, 2023c] propose Q-Former to align image representation and text representation, InstructBLIP [Dai *et al.*, 2023] based on BLIP-2 and propose an instruction tuning framework to improve the instruction following capability. CogVLM [Wang *et al.*, 2023] propose a visual expert module to enable deep alignment of the vision-language features. LLaVA [Liu *et al.*, 2023b; Liu *et al.*, 2023a], Emu2 [Sun *et al.*, 2023], and MiniGPT-4 [Zhu *et al.*, 2023] adopt a simple but effective projection scheme to connect image feature into the language space. A modality-adaptive module is introduced by mPLUG-Owl2 [Ye *et al.*, 2023], aiming to enhance modality collaboration by projecting visual and linguistic features into a shared space. In this paper, we will provide a comprehensive evaluation of some of these models using the CMMU benchmark and assess their abilities in domain-specific knowledge.

	Q	I	Exp	Question Type	Lang
MMLU	15,687	✓	×	MCQ	en
MMBench	2,974	✓	×	MCQ	en,ch
SCI.QA	21,208	✓	✓	MCQ	en
M3Exam	12,317	✓	×	MCQ	multilingual
M3KE	20,477	×	×	MCQ	zh
MME	2,374	✓	×	True or False	en
MMMU	11,500	✓	✓	MCQ	en
CMMU	3,603	✓	✓	MCQ, FBQ, MRQ	zh

Table 2: Compare CMMU with existing datasets. Q means quantity, I means image, Exp means explanation of the answer, Lang means language.

	Val Avg.	Test Avg.	Val-Normal MCQ	MRQ	FBQ	Val-Hard MCQ	MRQ	FBQ	Test-Normal MCQ	MRQ	FBQ	Test-Hard MCQ	MRQ	FBQ
InstructBLIP-13b	0.39	0.48	0.0	0.0	0.79	0.0	0.0	1.67	0.08	0.0	1.7	0.0	1.05	0.0
CogVLM-7b	5.55	4.9	5.98	0.0	6.9	2.0	2.13	5.0	5.89	0.0	5.1	0.67	0.0	4.73
ShareGPT4V-7b	7.95	7.63	8.71	0.0	9.27	7.33	1.06	6.0	8.38	0.0	10.4	2.67	0.0	5.41
mPLUG-Owl2-7b	8.69	8.58	10.62	3.03	8.28	6.67	1.06	5.67	9.63	0.0	11.15	5.33	1.05	4.73
LLava-1.5-13b	11.36	11.96	12.7	0.0	12.62	8.67	1.06	9.67	13.03	3.12	14.93	6.67	0.0	9.8
Qwen-VL-Chat-7b	11.71	12.14	9.71	3.03	17.36	3.33	1.06	18.67	10.62	0.0	21.36	0.67	1.05	12.5
Intern-XComposer-7b	17.87	18.42	22.49	3.03	16.96	8.67	4.26	11.33	22.16	12.5	20.04	7.33	1.05	12.16
Gemini-Pro	21.58	22.5	18.42	**24.24**	33.53	5.33	17.02	23.33	20.83	21.87	31.95	4.67	11.58	25.0
Qwen-VL-Plus	27.51	27.73	26.33	12.5	34.98	**19.46**	14.89	29.19	28.31	28.12	31.19	**22.82**	10.53	**27.12**
GPT-4V	**30.19**	**30.91**	**30.54**	21.21	**35.31**	14.67	**23.4**	**31.0**	**32.86**	**37.5**	**37.81**	12.67	**16.84**	23.65

Table 3: The accuracy of comparing models on different question types and difficulty levels. We report the results of the models on the validation and test sets.

3 The CMMU Benchmark

CMMU is a novel multi-modal benchmark designed to evaluate domain-specific knowledge across seven foundational subjects: math, biology, physics, chemistry, geography, politics, and history. We collect questions encompassing both text and images sourced from diverse exams. CMMU covers questions from primary school to high school, providing a comprehensive evaluation of the abilities of models across various grades.

Previous benchmarks, such as ScienceQA and MMMU, only have multiple-choice questions. In contrast, our CMMU benchmark contains 3 types of questions:

- Multiple-choice question (MCQ): Each question presents 3 or 4 options, with only one correct answer.
- Multiple-response question (MRQ): Each question includes 4 options, and the number of correct answers can range from 1 to 4.
- Fill-in-the-blank question (FBQ): The question is to fill in the blanks with the correct answers to complete the sentence or passage.

In addition to providing the correct answer, CMMU also provides the explanations of the answers about MCQ and MRQ.

3.1 Data Pre-process

Data Collection and Processing. We extract text and images from the original PDF and convert them into JSON format. In addition, we transform all formulas, including mathematical and chemical ones, into LaTeX format. For fill-in-the-blank questions, if there are many sub-questions within one question, we will split them into a sub-question list, attempting to have only one blank to fill in each sub-question, except for some indivisible questions. In the end, we decompose 639 fill-in-the-blank questions into 1,632 sub-questions, with 83% of them requiring only one blank to be filled.

Data Cleaning. We manually review the questions, filtering out images that are blurry, low-quality, or have a resolution less than 50×50 dpi, eliminating questions that are

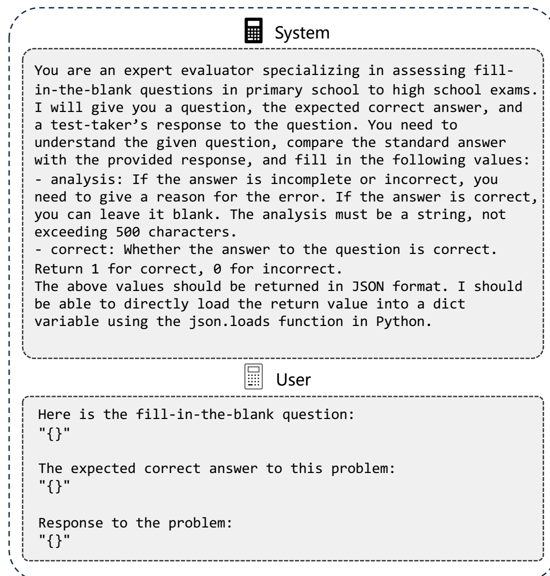


Figure 4: Prompt template used in fill-in-the-blank questions

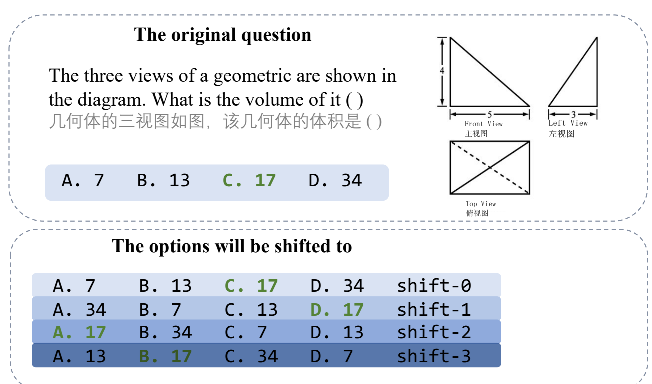


Figure 5: A demonstration of the Circular Evaluation in ShiftCheck.

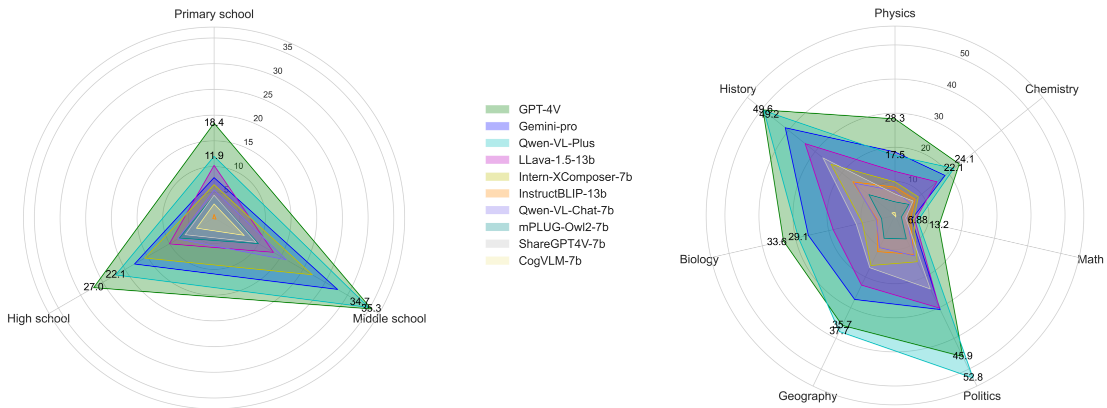


Figure 6: Overall results on the accuracy of different models in subjects and grades.

incorrectly parsed, and correcting mistakes made during the automatic conversion to LaTeX. Furthermore, experienced teachers consider the depth of knowledge and the complexity of question-solving methods to categorize each question into two levels: normal and hard.

3.2 Data Distribution

CMMU has a total of 3603 questions, divided into validation set and test set, with 1,800 and 1,803 questions respectively. The validation set will be open source to the community. As shown in Figure 2, the benchmark contains multi-modal content from middle and high school across 7 types of subjects, while primary school only contains math. The ratio of normal and hard questions are 8:2. Over 70% of the questions have detailed answer explanations, with an average length of 311 characters for each analysis. The detailed statistics are shown in Table 1. We translate the original questions into English and analyze the distribution of their first three words. As shown in Figure 3, the questions have diverse formats and are relevant to images or diagrams.

3.3 Comparisons with Existing Datasets

Table 2 shows the comparisons with existing datasets. We compare the benchmarks from five dimensions: quantity, with or without images, with or without explanation, question type and language. It can be seen that CMMU is the first benchmark for multi-modal and muli-type question understanding and reasoning in Chinese.

4 Evaluation

4.1 ShiftCheck Strategy

The evaluation of multiple-choice questions confronts two challenges: First, considering the particularity of the formats of multiple-choice questions, when the model correctly answers a question there is an uncertainty about whether the model has truly mastered the relevant knowledge or it just guesses the correct answer. When a model chooses answers through guessing, there will be positional bias, which means a LLM will prefer the answer in a certain position. Position bias is an issue that appears in many LLMs and MLLMs, however, existing methods have not quantitatively measured the extent of the position bias.

To address the above problems, we propose the ShiftCheck Strategy including two stages. Firstly, following the Circular Evaluation [Liu *et al.*, 2023c], we cyclically change the positions of the options and let the model answer questions. Subsequently, we calculate metrics to quantify position bias. We will describe the whole process in detail.

Circular Evaluation For a multiple-choice question with k options, we perform a right circular shift on the options. For example, if the original order of the options is $ABCD$, then after one shift, the order will change to $DABC$. A detailed example is provided in Figure 5. Given a question Q with k possible options, we generate k distinct shifted-option questions, denoted as Q_i, $i \in [0, k]$. Each Q_i is then evaluated by the MLLMs to generate the corresponding answers A_i. We consider the model to have sufficient knowledge to answer the question Q only if all of A_i are correct, in which case the accuracy score of Q is 1, otherwise it is 0.

Position Bias Measurement Conceptually, an unbiased model assigns equal probability to each option. Under the shifted-option setting, if the probability of each option is not equal, it indicates that the model has a bias towards a certain option. Considering this, we define the $BiasRate$ as follows:

Questions that are completely answered correctly in Circular Evaluation do not reflect position bias, so we just focus on the incorrectly answered questions. If there are M incorrectly answered questions with n options for each, there will be a total of $m*n$ answers combination. We count the occurrences S_o of each option o and then calculate the probability $P_o = \frac{S_o}{m*n}, o \in \{A, B, C, ...\}$. And then we define the BiasRate as the variance of P, the formula is $BiasRate = \sigma^2(P)$. The larger the $BiasRate$, the greater the positional bias of the model.

4.2 Evaluations on Different Question Types

To avoid the impact of the analysis process of model outputs on the evaluation, we filter the answers by retaining only the last line of the answer. For multiple-choice and multiple-response questions, we extract option letters from

	Physics	Chemistry	Math	Politics	Geography	Biology	History	Primary	Middle	High
InstructBLIP-13b	0.92	0.26	0.24	0.48	0.0	0.12	1.17	0.68	0.45	0.29
CogVLM-7b	4.0	5.23	1.93	7.66	7.41	4.05	9.79	2.73	6.74	3.95
ShareGPT4V-7b	7.23	5.75	3.49	13.16	10.47	5.64	15.15	4.44	9.44	6.5
Qwen-VL-Chat-7b	6.31	6.8	3.25	23.92	17.01	11.66	27.04	5.46	16.27	8.18
mPLUG-Owl2-7b	8.31	7.58	4.69	12.68	11.77	5.4	15.85	4.44	9.98	7.8
LLava-1.5-13b	9.85	8.76	5.66	15.07	16.28	9.82	24.01	10.24	13.35	10.06
Intern-XComposer-7b	12.92	15.82	6.02	30.14	22.67	18.65	33.8	6.48	22.07	15.59
Qwen-VL-Plus	17.57	22.12	6.88	**52.88**	**37.76**	29.15	49.29	11.99	34.79	22.14
Gemini-Pro	18.62	18.82	5.05	30.62	27.33	26.13	41.26	7.85	27.78	17.9
GPT-4V	**28.31**	**24.18**	**13.24**	45.93	35.76	**33.62**	**49.65**	**18.43**	**35.37**	**27.09**

Table 4: Detailed statistics of different models in subjects and grades. We average the accuracy of different difficulty questions and report the average values on the test and validation sets.

the responses of the models. After that, we apply different strategies to evaluate the three types of questions.

Evaluation on Multiple-choice Question: We adopt the ShiftCheck Strategy as mentioned in section 4.1, which allows us to analyze both the accuracy and the $BiasRate$.

Evaluation on Multiple-response Question: This question type may have more than one correct option. We consider the correctness only when all the chosen options are correct, excluding any incorrect choices.

Evaluation on Fill-in-the-blank Question: The answers to fill-in-the-blank questions may not be unique and responses with similar meanings to the groundtruth can also be considered correct. Hence, we utilize GPT-4 to judge the answer, providing a binary score of 0 or 1 to determine correctness. Further details about the evaluation prompts are in Figure 4.

5 Experiments

5.1 Models

We evaluate the performance of various MLLMs on the CMMU benchmark, including both closed-source and open-source models. The closed-source models are evaluated by using their official API, while open-source models are evaluated by running inferences on NVIDIA A100 GPUs. For the closed-source models, we select state-of-the-art models like GPT-4V, Gemini-Pro. We also choose Qwen-VL-Plus, which performs well on Chinese datasets. For the open-source models, model sizes vary from 7b to 13b, including LLava-1.5-13b, CogVLM-7b, InstructBLIP-13b, Qwen-VL-Chat-7b, Intern-XComposer-7b, mPLUG-Owl2-7b and ShareGPT4V-7b.

5.2 Prompts and Settings

All models are tested in zero-shot settings as we only specify the output format in prompts. Each type of question has its own prompt template, and we utilize the same prompt template for all models. The prompt[1] of MCQ is "Answer with the option's letter from the given choices directly", the prompt of MRQ is "Please directly provide the letters of the

[1]In the experiments, we use the Chinese version and we translate it into English for reading.

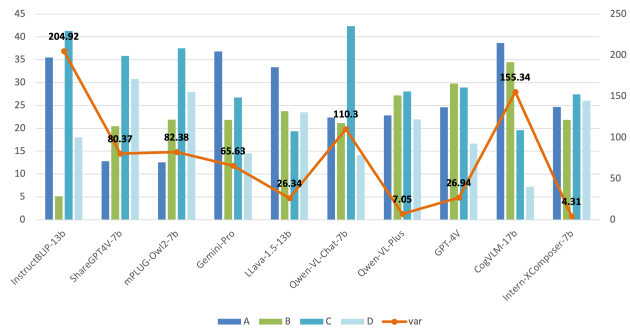


Figure 7: Option distribution and bias rate for different models

correct options. There may be more than one correct option." the prompt of FBQ is "Complete each blank with a single word or phrase. If there is more than one blank, split answers with a semicolon (;)".

Parameters are configured with distinct settings for each question type. For MCQ and MRQ, the temperature is set to 0, and the max new token is set to 10. For FBQ, the temperature is set to 0.2, and the max new token is set to 128.

5.3 Results

Results on Different Subjects and Grades

The overall results are shown in Table 3. GPT-4V achieves an accuracy of 30.19% and 30.91% on the validation and test sets, respectively, reaching the highest level among all models. Moreover, all three closed-source models perform better than all open-source models. From the perspective of question types, most models show lower accuracy in FBQ and MRQ compared to MCQ. Specifically, 7 models achieve an accuracy of less than 10% on the Val-hard set of MRQ, and 5 models achieve an accuracy of less than 10% on the Val-hard set of FBQ, highlighting the difficulty and challenge of FBQ and MRQ.

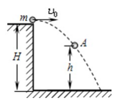
Results on Different Question Types

Figure 6 shows the overall results of different models in subjects and grades, and the detailed statistics are shown in Table 4. We can see that there is a significant subject bias

Q1: As shown in the diagram, there are two routes from point M to point N, denoted as A and B. Comparing the distances covered by these two routes, which of the following statements is true?
(A) Route A is longer.
(B) Route B is longer.
(C) Both are equally distant.
(D) Unable to determine.
Answer: C

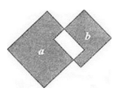
GPT4V Response:
Analyzing the question, there are two routes, A and B, from point M to point N. Route A is a straight line, while route B is a curve, and the length of the curve is noticeably longer than that of the straight line. Therefore, it can be concluded that route B is longer than route A.
The correct option is B.
Is correct : no
Mistake Type : image misinterpretation

Q2: As shown in the diagram, when an object of mass m leaves the tabletop with a velocity v_0 and reaches point A, the mechanical energy it possesses (with the tabletop as the reference for zero potential energy and neglecting air resistance) is ():
(A) $\frac{1}{2}mv_0^2$
(B) $\frac{1}{2}mv_0^2 + mgh$
(C) $\frac{1}{2}mv_0^2 - mgh$
(D) $\frac{1}{2}mv_0^2 + mg(H-h)$
Answer: A



GPT4V Response:
Analysis:
1. After the object leaves the tabletop and considering no air resistance, mechanical energy is conserved.
2. The object has initial kinetic energy $\frac{1}{2}mv_0^2$ on the tabletop with no potential energy (taking the tabletop as the reference for zero potential energy).
3. As the object passes through point A, kinetic energy decreases, and potential energy increases.
4. The increase in potential energy is mgh because the object has descended by a height h.
5. The decrease in kinetic energy equals the increase in potential energy. Therefore, the mechanical energy of the object at point A is the initial kinetic energy minus the increase in potential energy, which is $\frac{1}{2}mv_0^2 - mgh$
The correct option is C.
Is correct : no
Mistake Type : misleading reasoning

Q3: In the diagram, the areas of two squares are 25 and 16, and the shaded portions have areas a and b respectively. What is the value of $(a-b)$?
(A) 9
(B) 8
(C) 7
(D) 6
Answer: A



GPT4V Response:
Analysis:
1. From the given information, we know that the side lengths of the two squares are 5 and 4 respectively.
2. The shaded areas a and b are obtained by subtracting the area of the central uncovered square from a complete square.
3. For shaded area a, the area of the uncovered square is 16, so $a = 25 - 16 = 9$.
4. For shaded area b, the area of the uncovered square is 25, so $b = 16 - 25 = -9$.
5. Therefore, $(a - b) = 9 - (-9) = 18$.
6. However, there is no option for 18, indicating an error in our calculation.
7. Upon reevaluation, we realize that the central uncovered square is actually formed by the overlap of two squares, and its area is not 25 or 16, but the square of the difference between the side lengths of the two squares, i.e., $(5 - 4)^2 = 1^2 = 1$.
8. Therefore, the area of shaded part a should be $25 - 1 = 24$, and the area of shaded part b should be $16 - 1 = 15$.
9. Thus, $(a - b) = 24 - 15 = 9$.
The correct option is A.
Is correct : the answer is right but the analysis process is wrong
Mistake Type : question misunderstanding

Figure 8: Case study of GPT-4V. Green represents the correct answers to questions, red shows where errors occur in the reasoning process of the model, and blue provides the types of errors. The text is translated into English for reading.

in all models. Subjects such as politics and history rely on the knowledge reservoir have higher accuracy than subjects such as physics, math which require computation and reasoning. When comparing different grades, the results show that MLLMs generally perform better on middle school questions than high school ones, suggesting that more complex knowledge presents a more significant challenge for these models.

Position Bias Analysis
We employ the **ShiftCheck Strategy** for quantitative analysis of position bias. As shown in Figure 7, most models have a positional preference for one or two specific options. An interesting finding is that, although these positional preferences are inconsistent across models, none of them choose Option D as their most preferred choice. By analyzing the $BiasRate$, we find that superior models, such as GPT-4V, tend to have a relatively lower $BiasRate$.

Case Study with CoT Prompts
To further analyze the performance of models using Chain of Thought(CoT), we change the prompt of MCQ to "Please analyze the question step by step and eventually provide a single correct option letter. (This is a multiple-choice question.)" Then, we choose GPT-4V, which has a strong ability in instruction-following, to answer 500 randomly selected MCQs. We identify three common mistake types in the model outputs: image misunderstanding, misleading reasoning, and question misunderstanding, with proportions of 27.48%, 35.41%, and 13.03%, respectively. Cases in Figure 8 show the above common mistakes respectively: In Question 1, the model fails to identify the route A in the image correctly. In Question 2, the model thinks that the kinetic energy during free fall is transformed into gravitational potential energy, leading to an incorrect reasoning result. In Question 3, the model misunderstands the question and produces a hallucination that the overlap area is 1. Although it guesses the answer correctly, it cannot pass the ShiftCheck Strategy. All bad cases demonstrate that even one of the most advanced MLLMs cannot perfectly solve questions of CMMU, highlighting both the potential and challenges of the benchmark.

6 Conclusion and Future Work

In conclusion, our work introduces a novel benchmark named CMMU to evaluate the multi-modal and multi-type question understanding and reasoning abilities of MLLMs in Chinese. Unlike existing benchmarks focusing on multiple-choice questions, CMMU offers a more comprehensive evaluation by incorporating a broader question type, including MCQ, MRQ, and FBQ. We also propose the ShiftCheck approach to quantify the position bias of the model, minimize the impact of randomness and ensure evaluation correctness. The evaluation results contribute to a deeper understanding of current MLLMs in the context of diverse and complex question formats. In future work, we will consider enriching the problem types of CMMU and expanding the number of problems to increase the challenge of the benchmark further.

References

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

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

[Chiang *et al.*, 2023] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

[Dai *et al.*, 2023] W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2023.

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

[Goyal *et al.*, 2017] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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, 2017.

[Hudson and Manning, 2019] Drew A Hudson and Christopher D Manning. 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, 2019.

[Kafle and Kanan, 2017] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *ICCV*, 2017.

[Kazemzadeh *et al.*, 2014] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[Li *et al.*, 2023a] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023.

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

[Li *et al.*, 2023c] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 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, 2014.

[Liu *et al.*, 2023a] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[Liu *et al.*, 2023b] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[Liu *et al.*, 2023c] 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.

[Lu *et al.*, 2022] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[Macfarlane, 2023] Donald Macfarlane. Professional report generation using lexeme theories and openai's generative pretrained transformer, gpt-4: A comparison. *Medical Research Archives*, 11(11), 2023.

[Mialon *et al.*, 2023] Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*, 2023.

[OpenAI, 2023] R OpenAI. Gpt-4 technical report. *arXiv*, pages 2303–08774, 2023.

[Plummer *et al.*, 2015] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

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

[Sun *et al.*, 2023] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*, 2023.

[Team *et al.*, 2023] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Wang *et al.*, 2023] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

[Xu *et al.*, 2023] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023.

[Ye *et al.*, 2023] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023.

[Yue *et al.*, 2023] 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. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.

[Zhang *et al.*, 2023] Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *arXiv preprint arXiv:2306.05179*, 2023.

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

[Zhu *et al.*, 2016] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016.

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