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 an evaluation strategy called Positional Error Variance for assessing multiple-choice questions. The strategy aims to 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. The data and code are available at https://github.com/FlagOpen/CMMU.

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 shown 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 MM-Bench [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 multi-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 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, as shown in Table 1, 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 Positional Error Variance approach to measure the position bias for multiple-choice question. Specifically, we cycle through the position of options to ensure that the answer can appear at any position with equal probability, which is the same as CircularEval, aim at reducing position bias, minimizing the influence of randomness on correctness. Then we produce Positional Error Variance, a quantitative analysis to measure 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 Positional Error Variance, which is designed to conduct a quantitative analysis of position bias in MLLMs.

### 2 Related Work

#### 2.1 Multi-modal Benchmarks

With the development of large language models (LLMs), there is a growing emphasis in research communities on assessing the capabilities of LLM such as HELM [Liang et al., 2022], CLEVA [Li et al., 2023d], especially their multimodal understanding abilities. Datasets like VQA2 [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 ReferCOCO [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 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. MM-Bench [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. BenchLMM [Cai et al., 2023] assesses MLLMs in three distinct styles: artistic style, sensor style, and application style. 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., 2024] 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. LLaV A [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.

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
The three views of a geometric are shown in the diagram. What is the volume of it?

The original question:

A. 7  B. 13  C. 17  D. 34

The options will be shifted to:

A. 7  B. 13  C. 17  D. 34  shift-0
A. 34  B. 7  C. 13  D. 17  shift-1
A. 17  B. 34  C. 7  D. 13  shift-2
A. 13  B. 17  C. 34  D. 7  shift-3

Figure 3: A demonstration of the CircularEval.
answer explanations, with an average length of 311 characters for each analysis. The detailed statistics are shown in Table 2. We translate the original questions into English and analyze the distribution of their first three words. As shown in Figure 2, the questions have diverse formats and are relevant to images or diagrams.

### 3.3 Comparisons with Existing Datasets

Table 3 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 multi-type question understanding and reasoning in Chinese.

### 4 Evaluation

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 Positional Error Variance. Firstly, following the CircularEval \[Liu \textit{et al.}, 2023c\], 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.

#### 4.1 CircularEval

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 3. 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.

#### 4.2 Positional Error Variance

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 \( \text{BiasRate} \) as follows:

Questions that are completely answered correctly in CircularEval 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 \( m \times n \) answers combination. We count the occurrences \( S_o \) of each option \( o \) and then calculate the probability \( P_o = \frac{S_o}{m \times n}, o \in \{A, B, C, \ldots\} \). And then we define the BiasRate as the variance of \( P \), the formula is \( \text{BiasRate} = \sigma^2(P) \). The larger the \( \text{BiasRate} \), the greater the positional bias of the model.

#### 4.3 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 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 CircularEval and Positional Error Variance in section 4, which allows us to analyze both the accuracy and the \( \text{BiasRate} \).

**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.
Table 6: 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Physics</th>
<th>Chemistry</th>
<th>Math</th>
<th>Politics</th>
<th>Geography</th>
<th>Biology</th>
<th>History</th>
<th>Primary</th>
<th>Middle</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>InstructBLIP-13b</td>
<td>0.92</td>
<td>0.26</td>
<td>0.24</td>
<td>0.48</td>
<td>0.0</td>
<td>0.12</td>
<td>1.17</td>
<td>0.68</td>
<td>0.45</td>
<td>0.29</td>
</tr>
<tr>
<td>CogVLM-7b</td>
<td>4.0</td>
<td>5.23</td>
<td>1.93</td>
<td>7.66</td>
<td>7.41</td>
<td>4.05</td>
<td>9.79</td>
<td>2.73</td>
<td>6.74</td>
<td>3.95</td>
</tr>
<tr>
<td>ShareGPT4V-7b</td>
<td>7.23</td>
<td>5.75</td>
<td>3.49</td>
<td>13.16</td>
<td>10.47</td>
<td>5.64</td>
<td>15.15</td>
<td>4.44</td>
<td>9.44</td>
<td>6.5</td>
</tr>
<tr>
<td>Qwen-VL-Chat-7b</td>
<td>6.31</td>
<td>6.8</td>
<td>3.25</td>
<td>23.92</td>
<td>17.01</td>
<td>11.66</td>
<td>27.04</td>
<td>5.46</td>
<td>16.27</td>
<td>8.18</td>
</tr>
<tr>
<td>mPLUG-Owl2-7b</td>
<td>8.31</td>
<td>7.58</td>
<td>4.69</td>
<td>12.68</td>
<td>11.77</td>
<td>5.4</td>
<td>15.85</td>
<td>4.44</td>
<td>9.98</td>
<td>7.8</td>
</tr>
<tr>
<td>LLava-1.5-13b</td>
<td>9.85</td>
<td>8.76</td>
<td>5.66</td>
<td>15.07</td>
<td>16.28</td>
<td>9.82</td>
<td>24.01</td>
<td>10.24</td>
<td>13.35</td>
<td>10.06</td>
</tr>
<tr>
<td>Intern-XComposer-7b</td>
<td>12.92</td>
<td>15.82</td>
<td>6.02</td>
<td>30.14</td>
<td>22.67</td>
<td>18.65</td>
<td>33.8</td>
<td>6.48</td>
<td>22.07</td>
<td>15.59</td>
</tr>
<tr>
<td>Qwen-VL-Plus</td>
<td>17.57</td>
<td>22.12</td>
<td>6.88</td>
<td>32.88</td>
<td>37.76</td>
<td>29.15</td>
<td>49.29</td>
<td>11.99</td>
<td>34.79</td>
<td>22.14</td>
</tr>
<tr>
<td>Gemini-Pro</td>
<td>18.62</td>
<td>18.82</td>
<td>5.05</td>
<td>30.62</td>
<td>27.33</td>
<td>26.13</td>
<td>41.26</td>
<td>7.85</td>
<td>27.78</td>
<td>17.9</td>
</tr>
<tr>
<td>GPT-4V</td>
<td>28.31</td>
<td>24.18</td>
<td>13.24</td>
<td>45.93</td>
<td>35.76</td>
<td>33.62</td>
<td>49.65</td>
<td>18.43</td>
<td>35.37</td>
<td>27.09</td>
</tr>
</tbody>
</table>

5 Experiments

5.1 Models

We evaluate the performance of various MLLMs, 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 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 4. 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 4 shows the overall results of different models in subjects and grades, and the detailed statistics are shown in Table 6. We can see that there is a significant subject bias 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 perform better on middle school questions than high school ones, suggesting that more complex knowledge presents a more significant challenge for these models.

\(^1\)In the experiments, we use the Chinese version and we translate it into English for reading.
**Position Bias Analysis**

We employ the **Positional Error Variance** for quantitative analysis of position bias. As shown in Figure 5, 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 Table 7 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 CircularEval. 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 Positional Error Variance to quantify the position bias of MLLMs. 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 and increasing the quantity to increase the challenge of the benchmark further.

**Acknowledgments**

This work is supported by National Key R&D Program of China (2022ZD0116306). Thanks to Duo Zheng and Ning Su for their valuable suggestions and contributions to this paper.

**Contribution Statement**

Zheqi He and Xinya Wu contributed equally to this work. This research was conducted during Pengfei Zhou’s internship at the Beijing Academy of Artificial Intelligence. Guang Liu and Xi Yang serve as corresponding authors for this study.

---

**Table 7: 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.**

![Image](https://via.placeholder.com/150)

**Question and Response**

<table>
<thead>
<tr>
<th>Question</th>
<th>Analysis and Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Q1:</strong> 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?</td>
<td><strong>GPT4V Response:</strong> 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 answer is B. <strong>Is correct:</strong> no <strong>Mistake Type:</strong> image misinterpretation</td>
</tr>
<tr>
<td>(A) Route A is longer.</td>
<td><strong>Is correct:</strong> no <strong>Mistake Type:</strong> misleading reasoning</td>
</tr>
<tr>
<td>(B) Route B is longer.</td>
<td><strong>Is correct:</strong> no <strong>Mistake Type:</strong> question misunderstanding</td>
</tr>
<tr>
<td>(C) Both are equally distant.</td>
<td><strong>Is correct:</strong> no <strong>Mistake Type:</strong> question misunderstanding</td>
</tr>
<tr>
<td>(D) Unable to determine.</td>
<td><strong>Is correct:</strong> no <strong>Mistake Type:</strong> question misunderstanding</td>
</tr>
<tr>
<td><strong>Answer:</strong> C</td>
<td><strong>GPT4V Response:</strong> 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 answer is B. <strong>Is correct:</strong> no <strong>Mistake Type:</strong> image misinterpretation</td>
</tr>
<tr>
<td><strong>Q2:</strong> As shown in the diagram, when an object of mass m reaches the top of the table with a velocity ( v_0 ) and reaches point A, the mechanical energy it possesses (with the table as the reference for zero potential energy and neglecting air resistance) is ( ):</td>
<td><strong>GPT4V Response:</strong> 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 answer is B. <strong>Is correct:</strong> no <strong>Mistake Type:</strong> image misinterpretation</td>
</tr>
<tr>
<td>(A) ( \frac{1}{2}mv_0^2 )</td>
<td><strong>Is correct:</strong> no <strong>Mistake Type:</strong> misleading reasoning</td>
</tr>
<tr>
<td>(B) ( \frac{1}{2}mv_0^2 + mgh )</td>
<td><strong>Is correct:</strong> no <strong>Mistake Type:</strong> question misunderstanding</td>
</tr>
<tr>
<td>(C) ( \frac{1}{2}mv_0^2 + mghh )</td>
<td><strong>Is correct:</strong> no <strong>Mistake Type:</strong> question misunderstanding</td>
</tr>
<tr>
<td>(D) ( \frac{1}{2}mv_0^2 + mghH )</td>
<td><strong>Is correct:</strong> no <strong>Mistake Type:</strong> question misunderstanding</td>
</tr>
<tr>
<td><strong>Answer:</strong> A</td>
<td><strong>GPT4V Response:</strong> 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 answer is B. <strong>Is correct:</strong> no <strong>Mistake Type:</strong> image misinterpretation</td>
</tr>
<tr>
<td><strong>Q3:</strong> 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 )?</td>
<td><strong>GPT4V Response:</strong> 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 answer is B. <strong>Is correct:</strong> no <strong>Mistake Type:</strong> image misinterpretation</td>
</tr>
<tr>
<td>(A) 9</td>
<td><strong>Is correct:</strong> no <strong>Mistake Type:</strong> misleading reasoning</td>
</tr>
<tr>
<td>(B) 8</td>
<td><strong>Is correct:</strong> no <strong>Mistake Type:</strong> question misunderstanding</td>
</tr>
<tr>
<td>(C) 7</td>
<td><strong>Is correct:</strong> no <strong>Mistake Type:</strong> question misunderstanding</td>
</tr>
<tr>
<td>(D) 6</td>
<td><strong>Is correct:</strong> no <strong>Mistake Type:</strong> question misunderstanding</td>
</tr>
<tr>
<td><strong>Answer:</strong> A</td>
<td><strong>GPT4V Response:</strong> 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 answer is B. <strong>Is correct:</strong> no <strong>Mistake Type:</strong> image misinterpretation</td>
</tr>
</tbody>
</table>
References


