

Supplementary Material for M4U: Evaluating Multilingual Understanding and Reasoning for Large Multimodal Models

A. Dataset Documentation

A.1. Usage

- M4U and M4U-mini dataset are public available at <https://huggingface.co/datasets/M4U-Benchmark/M4U> and <https://huggingface.co/datasets/M4U-Benchmark/M4U-mini>, respectively. We provide the evaluation code for LLaVA and GPT-4o at <https://github.com/M4U-Benchmark/M4U>.
- The dataset is saved in Parquet format. We present the data format and the examples in the README.md file. Besides, we also provide the example codes to show how to evaluate the multimodal models on the dataset.
- License: M4U is under CC BY-NC-SA Liences. The guidelines for annotators stress the importance of strictly following copyright and licensing rules from the original data sources, particularly avoiding materials from websites that prohibit copying and redistribution. If you come across any data samples that may violate copyright or licensing regulations, please inform us. Once verified, such samples will be promptly removed.

A.2. Data sources

M4U consists of 10,005 multiple-choices questions, covering 64 disciplines of 16 subfields from Science, Engineering and Healthcare. To minimize the risk of data contamination, the samples are collected from college exams, the quizzes of online video lectures. Further a large portion (35%) of the questions are written by our team according to the textbooks. The guidelines for annotators stress the importance of strictly following copyright and licensing rules from the original data sources, particularly avoiding materials from websites that prohibit copying and redistribution.

A.3. Evaluation prompt

We provide the prompt template used for the zero-shot evaluation in the following part. For chain-of-thought prompting, the model is first instructed to generate a rationale based on the question and answer options. This rationale is then appended after the options. Finally, the model is prompted to follow a specific instruction to directly generate the predicted

answer choice for each question.

Evaluation template to extract rationales

```
{Question}
A. {1st Option}
B. {2nd Option}
C. {3rd Option}
D. {4th Option}
Please analyze the question and options.
```

Evaluation template for English

```
{Question}
A. {1st Option}
B. {2nd Option}
C. {3rd Option}
D. {4th Option}
{Generated Chain-of-Thought} (Optional)
Answer with the option's letter from the given
choices directly.
```

A.4. First-stage Processing Guidelines

We summarize the detailed first-stage processing guidelines for the annotators. For each dimension, we require the annotator to score the sample following the below guidelines. The question with the higher scores indicates higher quality. We filter out the questions with average scores lower than 2.0. As for the image quality, the standard is:

- 0 score: The image is extremely blurry, difficult to recognize, or most of it is cropped, resulting in severe information loss.
- 1 score: The image is relatively blurry, details are hard to discern, or parts of the image are cropped, leading to some information loss.
- 2 score: The image is slightly blurry; most content is recognizable but details are unclear, or the image is slightly cropped, but most information is complete.
- 3 score: The image is mostly clear; all major content is recognizable, though some details may not be clear.
- 4 score: The image is clear; all content and details are easily recognizable with no apparent defects.

- 5 score: The image is very clear; details are excellently represented, complete without any cropping or obstructions, meeting or exceeding the expected quality standards.

The guideline for measuring the question description quality is:

- 0 score: The question is vague and completely unintelligible, with no clear intent.
- 1 score: The question statement is ambiguous, difficult to fully understand its intent, with multiple possible interpretations.
- 2 score: The question statement is basically clear, but there are some ambiguities or lack of rigor that need further clarification.
- 3 score: The question statement is clear, though there are some details that are not rigorous or there is slight ambiguity.
- 4 score: The question statement is both clear and rigorous, with details well handled, and only very minor issues present.
- 5 score: The question statement is extremely clear and rigorous, logical, without any ambiguity, fully meeting high standards.

The standard for measuring the difficulty of visual understanding is:

- 0 score: The question almost does not rely on visual ability, can be fully understood without any visual information.
- 1 score: The question does not completely rely on visual ability, both visual and non-visual information are balanced.
- 2 score: Although the question relies on visual ability, a considerable proportion of non-visual information assists understanding.
- 3 score: The question largely depends on visual ability, but some non-visual information is provided.
- 4 score: The question greatly depends on visual ability, with very little content provided by non-visual information.
- 5 score: The question completely depends on visual ability, without it, the content is incomprehensible.

Furthermore, we recruit a team of graduate student of related majors to assess the difficulty and correctness for the questions. The team will filter out the questions with wrong answer or minor visual efforts. We present the distribution of image resolution for M4U dataset in Figure 1.

B. More Results

Different Image Types. We demonstrate the visualization of the detailed results of various LMMs on different image types in Figure 2. We reclassified 13 image types into 7 categories based on the style and application of the image. As shown in the top part of Figure 2, GPT-4o shows impressive performance on the image type of Plots & Charts and Medical compared with the other models, but has unsatisfac-

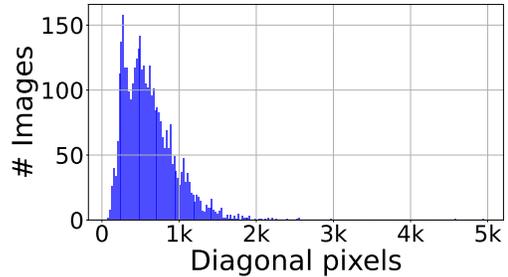


Figure 1. The distribution of image resolution for M4U dataset.

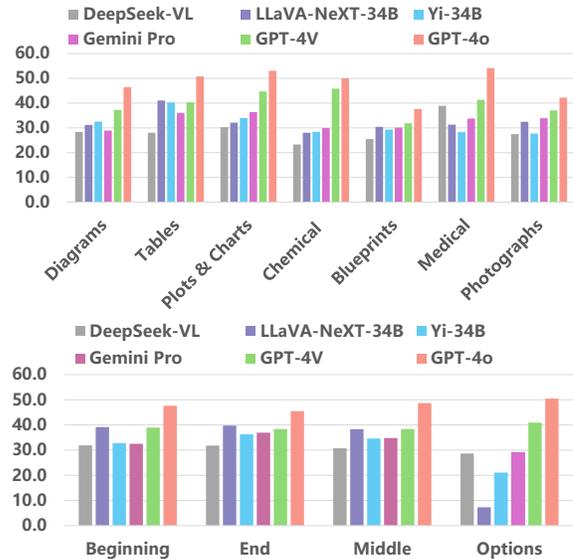


Figure 2. The zero-shot accuracy of different LMMs on different image types (Top) and positions (Bottom) on M4U dataset.

tory results on Blueprints. We argue that this is because the Blueprints contain many engineering sketches that require the capability of the fine-grained perception and domain-specific knowledge about engineering standards. We present more results about different subjects, disciplines and languages in the supplementary material.

Different Image Positions. M4U not only supports the image-text pairs as the input, but includes interleaved image-text documents. Thus, we conduct the analysis about the performance of the selected LMMs on different positions of the images. We divide these questions into four groups according to the image position: image at the beginning, end, middle of the question and in the options. As shown in the bottom part of Figure 2, on the questions with images in the options, GPT-4o and GPT-4V outperform the other models by a large gain, and LLaVA-NeXT-34B performs poorly on this types of the questions. We argue that this is because the LLaVA-NeXT series are only trained with a

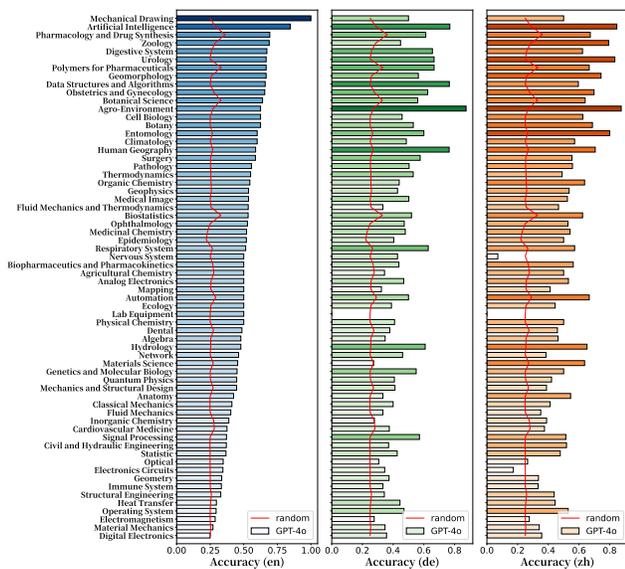


Figure 3. The zero-shot accuracy of GPT-4o across 64 subjects on M4U dataset.

high-quality corpus of image-text pairs. Instead DeepSeek-VL is pre-trained with a large mixture of image-text pairs and interleaved documents, and it does not suffer from a significant degradation of performance on the questions with images in the options.

Different Subjects, Disciplines and Languages. We present the detailed results of various LMMs on different fields of Chinese, English and German in Table 1. GPT-4o outperforms the other models by large improvements on all fields of all languages. For the open-source models, we observe that LLaVA-NeXT-34B shows impressive results on scientific reasoning, and DeepSeek-VL demonstrates good performance on Health. Further, we observe that on Science, the chain-of-thought prompting significantly improves the performance of GPT-4V by a gain of over 6% accuracy in Chinese and English, while only boosts the performance by an improvement of 3.1% accuracy on German. The similar phenomenon also exists for Gemini 1.0 Pro. On Health part, Gemini 1.0 Pro with the chain-of-thought prompting outperforms it with zero-shot prompting by a gain of 4.9% and 4.7% on English and German, but it leads to a degradation of 1.0% accuracy on Chinese. These results show that the effect of the chain-of-thought prompting also differ from different languages. Furthermore, we provide the detailed results of GPT-4o across 64 subjects on the test set of M4U in the supplementary material. We illustrate the zero-shot accuracy of GPT-4o across 64 subjects on English, German and Chinese set of M4U dataset in Figure 3.

Furthermore, we present a representative failure case of the translation-then-questioning approach in Figure 4. In

Question: <image> 长沙的晚秋のある日午後 \$20:00\$ に測定された気温は \$12^{\circ}\text{C}\$、湿球温度は \$10^{\circ}\text{C}\$ です。当日の夜、長沙は月明かりがあり星がまばらに見えました。長沙の翌朝に地面の霜が発生するかどうかを尋ねます。露点はどれくらいですか？指定されたオプションの文字を使って直接回答してください。

<image> On a late autumn day in Changsha, the temperature measured at 8:00 PM was \$12^{\circ}\text{C}\$, and the wet-bulb temperature was \$10^{\circ}\text{C}\$. That night, Changsha had moonlight and sparse stars visible. Will frost form on the ground in Changsha the next morning? What is the dew point? Please answer directly using the letter provided from the options.

Options:

A. ある, \$7.9^{\circ}\text{C}\$ Yes, \$7.9^{\circ}\text{C}\$ B. ある, \$-7.9^{\circ}\text{C}\$ Yes, \$-7.9^{\circ}\text{C}\$
 C. ない, \$7.9^{\circ}\text{C}\$ No, \$7.9^{\circ}\text{C}\$ D. ない, \$-7.9^{\circ}\text{C}\$ No, \$-7.9^{\circ}\text{C}\$

Translated Question: <image> 长沙的秋季的一天晚上8点测得气温为 \$12^{\circ}\text{C}\$，露点温度为 \$10^{\circ}\text{C}\$。当天晚上，长沙可以看到月亮，星星也十分明亮。长沙的早晨地表会结霜吗？露点是多少？直接用给定选项的字母回答。

Options :

A.会, \$7.9^{\circ}\text{C}\$ Yes, \$7.9^{\circ}\text{C}\$ B.会, \$-7.9^{\circ}\text{C}\$ Yes, \$-7.9^{\circ}\text{C}\$
 C.不会, \$7.9^{\circ}\text{C}\$ No, \$7.9^{\circ}\text{C}\$ D.不会, \$-7.9^{\circ}\text{C}\$ No, \$-7.9^{\circ}\text{C}\$

附：饱和水汽压表

t(\$^{\circ}\text{C}\$)	7.5	7.9	8.0	8.5	9.0	9.5	10.0	10.5	11.0	11.5	12.0
E(hp)	10.4	10.6	10.7	11.1	11.5	11.9	12.3	12.7	13.1	13.6	14.0

Answer : Ground truth: A End-to-end: A Translate-then-questioning: C

Reason: During the translation, Gemini 1.5 Flash mistakenly translated the wet-bulb temperature as the dew point temperature. The wet-bulb temperature is the adiabatic saturation temperature reached when air comes into direct contact with water, achieving a stable heat-moisture balance under constant pressure. The dew point temperature is the temperature at which air cools to the saturation point without changing the water vapor content and air pressure. For unsaturated air, the dew point temperature is less than the wet-bulb temperature; for saturated air, the dew point temperature equals the wet-bulb temperature.

Figure 4. A typical failed case of two-stage translate-then-questioning method.

this example, Gemini 1.5 Flash mistranslates “wet-bulb temperature” as “dew point temperature,” leading to an incorrect answer.

C. Qualitative Analysis

We conduct a qualitative analysis of GPT-4V under chain-of-thought prompting. Specifically, we randomly sample 75 questions (2.5%) from various disciplines across each language. These samples are selected where GPT-4V exhibits errors in its response or reasoning in at least one language. We categorize the causes of these errors into six types: perceptual error, lack of knowledge, reasoning error, textual understanding, annotation error, and answer extraction error.

The distribution of error categories is illustrated in Figure 5. Perceptual errors, lack of knowledge, and reasoning errors account for the majority of failure cases—comprising 96% in Chinese, 95% in English, and 92% in German. Notably, GPT-4V tends to show more knowledge-related errors in Chinese, while reasoning errors are more frequent in German and English. Qualitative examples of each error type are provided in the supplementary material.

Perceptual error is the most frequent error made by GPT-4V. It corresponds to the illusion phenomenon that occurs when extracting visual information from images provided by the questions. These kinds of hallucination could be divided in two main categories: visual information deficiency and misinterpretation. Figure 13 shows a typical case for the visual information misinterpretation: the extracted informa-

Table 1. The detailed results of different LMMs on Health, Science and Engineering of M4U dataset. Sci. and Eng. are short for Science and Engineering, respectively.

Models	Chinese			English			German		
	Health	Sci.	Eng.	Health	Sci.	Eng.	Health	Sci.	Eng.
Yi-VL-6B	31.2	34.1	34.9	32.1	32.2	30.0	29.0	29.2	30.8
DeepSeek-VL	40.1	22.6	28.5	38.0	31.9	28.6	<u>35.2</u>	29.3	27.8
Yi-34B	32.9	34.1	33.6	34.0	<u>33.2</u>	<u>32.6</u>	29.4	<u>30.2</u>	<u>32.0</u>
LLaVA-NeXT-34B	<u>38.1</u>	40.4	37.0	<u>37.2</u>	36.8	34.7	36.9	34.2	34.5
Gemini 1.0 Pro	38.8	34.5	31.4	34.9	33.4	29.8	33.1	30.5	28.8
+ Chain-of-thought prompting	37.8	33.3	32.6	38.8	33.3	30.6	37.8	32.2	31.8
GPT-4V(ision)	41.9	39.3	37.9	43.9	37.8	36.6	41.1	36.0	34.6
+ Chain-of-thought prompting	<u>43.9</u>	<u>46.2</u>	<u>41.7</u>	<u>45.8</u>	<u>43.3</u>	<u>41.9</u>	<u>42.5</u>	<u>39.1</u>	<u>39.3</u>
GPT-4o	56.0	47.3	45.0	56.2	44.3	42.8	52.9	40.9	43.0

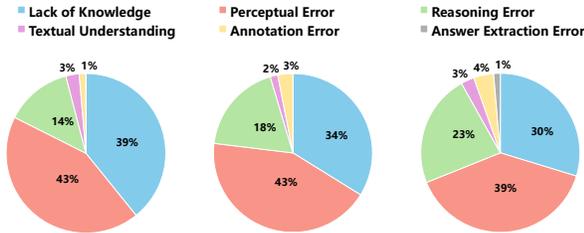


Figure 5. The distribution of the wrong cases of GPT-4V in different languages (Left: Chinese, Middle: English, Right: German).

tion is complete but contains mistakes. A portion of these mistakes are common perceptual errors in OCR and visual localization.

We define the **lack of knowledge** as the model has factual misunderstanding about the key concepts in questions and provides erroneous premise to the reasoning process. As shown in Figure 22, GPT-4V equates the average kinetic energy of a molecule to the kinetic energy of a single molecule, overlooking key preconditions of physical laws.

The **reasoning error** is categorized as the mathematical miscalculations and logical errors in the analysis procedure, which often occur in subjects need numerical computations and logical inference, such as math, physics, and electronics. As demonstrated in Figure 23, GPT-4V only considers the power supply on the left and does not consider the power supply on the right.

Remaining error cases (**others**) only occupy a small portion in selected cases, while depict long-tailed but various error reasons including textual misunderstanding, annotation error, and answer extraction error. Annotation error caused by typo or translation issues maintains less than 5% after manually checked by annotators.

Agriculture: Agroecology

CHINESE

Question: {answer}图中生物处于第一营养级?

Options:

A. "<image_1>"

B. "<image_2>"

C. "<image_3>"

D. "<image_4>"

<image_1>

<image_2>

<image_3>

<image_4>

GPT-4V: 根据题目, 需要找出图像中属于第一营养级的生物。第一营养级通常是指生态系统中的生产者, 它们通过光合作用或化能合成自养营养, 通常是植物。现在, 我们来分析每个选项: A选项: 这张图像显示了一片植被, 看上去像是一种草。这代表了植物, 属于生态链的第一营养级。B选项: 图像中是两只老鼠。老鼠是食草或杂食动物, 属于消费者, 不是第一营养级。C选项: 图像显示了两头狼。狼是肉食动物, 也是消费者, 不属于第一营养级。D选项: 最后这张图是一头牛。牛虽然是食草动物, 但作为消费者, 牛吃草, 因此不属于第一营养级。所以, 正确答案应该是选项A, 因为图像中的植物是属于第一营养级的生物。

GPT-4V (Answer): A.

Ground Truth: A.

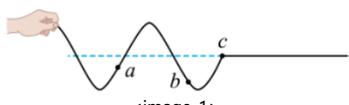
Figure 6. A sample correct case (subject: agriculture, language: Chinese).

Physics: Classical Mechanics **GERMAN**

Question: Nehmen Sie ein langes flexibles Seil, halten Sie ein Ende fest und ziehen Sie es gerade. Wenn Sie es dann kontinuierlich periodisch nach oben und unten schwingen lassen, können Sie eine Reihe von Wellen auf dem Seil sehen, wobei a, b, c Massenpunkte auf dem Seil sind. Im Moment, wenn die Welle gerade den Punkt c erreicht hat, ist die Wellenform auf dem Seil wie in der Abbildung <image_1> dargestellt. Welche der folgenden Aussagen ist korrekt?

Options:

- A. Die Hand beginnt sich nach oben zu bewegen
- B. Danach kehrt der Punkt a vor dem Punkt b zur Gleichgewichtslage zurück
- C. Zu diesem Zeitpunkt ist die Geschwindigkeit des Punktes a größer als die Geschwindigkeit des Punktes b
- D. Die Wellenform im Bild entspricht dem Zustand 1,5 Perioden nach Beginn der Schwingung der Wellenquelle



<image_1>

GPT-4V: Um die Frage zu beantworten, betrachten wir zunächst das Bild und die dargestellte Wellenform. Wir sehen einen Momentaufnahme des Seils, das eine stehende Welle ausbildet. Die Punkte a, b, und c sind Markierungspunkte auf dem Seil. Jetzt analysieren wir die Optionen: A. Die Hand beginnt sich nach oben zu bewegen. Ohne zusätzlichen Kontext ist es schwierig zu sagen, ob diese Aussage korrekt ist. Wir können aus dem Bild nicht direkt ableiten, in welche Richtung sich die Hand bewegt. Diese Aussage könnte man nur bestätigen, wenn man Informationen darüber hätte, welche Bewegung vor diesem Moment stattgefunden hat. B. Danach kehrt der Punkt a vor dem Punkt b zur Gleichgewichtslage zurück. Punkt b ist auf der Welle höher als Punkt a und wird daher eine längere Strecke zurücklegen, um zur Gleichgewichtslage zurückzukehren, als Punkt a. Das macht diese Aussage plausibel, aber wir brauchen mehr Kontext, um definitiv zu sein. C. Zu diesem Zeitpunkt ist die Geschwindigkeit des Punktes a größer als die Geschwindigkeit des Punktes b. Diese Aussage ist abhängig von der Position im Wellenzyklus. Da Punkt a näher an der Gleichgewichtslage ist, wo die Geschwindigkeit der schwingenden Punkte am größten ist, während Punkt b sich am Scheitelpunkt (einem Punkt maximaler Auslenkung der Welle) befindet, wo die Geschwindigkeit momentan null ist, ist diese Aussage wahrscheinlich korrekt. D. Die Wellenform im Bild entspricht dem Zustand 1,5 Perioden nach Beginn der Schwingung der Wellenquelle. Da wir nur ein Bild des Seils haben und keine Information über die Zeit, die vergangen ist, können wir nicht genau sagen, welcher Zustand dies in Bezug auf die Perioden der Welle darstellt. Es ist unmöglich, diese Aussage anhand des vorliegenden Bildes zu überprüfen. Nach dieser Analyse scheint Option C die plausibelste Aussage zu sein, basierend auf dem, was wir über die Dynamik von Wellen und Bewegungen von Punkten auf diesen Wellen wissen. In dem Moment, da Punkt b seinen Höhepunkt erreicht hat und seine Geschwindigkeit gleich null ist, müsste die Geschwindigkeit von Punkt a, der sich in der Nähe der Gleichgewichtslage befindet, größer sein. Antwort: C

GPT-4V (Answer) : C.

Ground Truth: C.

Figure 7. A sample correct case (subject: physics, language: German).

Agriculture: Agroecology **ENGLISH**

Question: The organism in {answer} is at the first trophic level?

Options:

- A. "<image_1>"
- B. "<image_2>"
- C. "<image_3>"
- D. "<image_4>"



<image_1> <image_2> <image_3> <image_4>

GPT-4V: The first trophic level in an ecological system is made up of producers; these are typically plants or algae that produce their own food through photosynthesis. From the images you've provided, we have: A. Grass in a field. B. A cow. C. Two wolves. D. Two rats. The organism at the first trophic level here would be the grass in image A, as it is capable of photosynthesis and can produce its own food from sunlight. The cow, wolves, and rats are all consumers at various trophic levels: The cow (B) is a primary consumer, as it eats plants. The wolves (C) are secondary or tertiary consumers, as they would typically eat other animals. The rats (D) can be primary consumers if they eat plants, or secondary if they eat other animals or omnivores if they consume both. So, the correct answer to your question is: A. "<image_1>", which shows grass at the first trophic level.

GPT-4V (Answer) : A.

Ground Truth: A.

Figure 8. A sample correct case (subject: agriculture, language: English).

Agriculture: Agroecology **GERMAN**

Question: {answer} im Bild befindet sich auf der ersten trophischen Ebene?

Options:

- A. "<image_1>"
- B. "<image_2>"
- C. "<image_3>"
- D. "<image_4>"



<image_1> <image_2> <image_3> <image_4>

GPT-4V: In einem Ökosystem sind die verschiedenen trophischen Ebenen die Stufen der Nahrungskette, beginnend mit den Produzenten an der Basis, gefolgt von den verschiedenen Konsumenten. Hier eine kurze Erläuterung jeder trophischen Ebene: Die erste trophische Ebene wird von Primärproduzenten besetzt, das sind in der Regel Pflanzen oder Algen, die durch Photosynthese Energie aus Sonnenlicht in chemische Energie umwandeln. Die zweite trophische Ebene umfasst Primärkonsumenten oder Herbivore, die sich von den Primärproduzenten ernähren. Die dritte Ebene besteht aus Sekundärkonsumenten, typischerweise Fleischfresser, die sich von Herbivoren ernähren. Darüber hinaus gibt es weitere Ebenen mit Tertiärkonsumenten und so weiter, die sich von Sekundärkonsumenten ernähren. Von den Bildern, die Sie hochgeladen haben, wäre das Bild, das die Organismen der ersten trophischen Ebene zeigt (Primärproduzenten), das mit Pflanzen darauf, da sie die Energie direkt von der Sonne durch Photosynthese nutzen. Also, um Ihre Frage zu beantworten: A. "<image_1>": Pflanzen, gehören zu den Primärproduzenten. B. "<image_2>": Ratten, gehören zu den Konsumenten. C. "<image_3>": Wölfe, gehören zu den Konsumenten. D. "<image_4>": Eine Kuh, gehört zu den Konsumenten (Herbivoren). Die erste trophische Ebene im Bild befindet sich auf: A. "<image_1>", da dies das Bild mit den Pflanzen ist.

GPT-4V (Answer) : A.

Ground Truth: A.

Figure 9. A sample correct case (subject: agriculture, language: German).

Geography: Geophysics **CHINESE**

Question: 判断正误, 日环食发生时, 月球、地球和太阳三者间的位置关系正确的是 <image_1> :

Options:
A. 对
B. 错



<image_1>

GPT-4V: 这张图片描绘了太阳、地球和月球的相对位置。日环食发生时, 月球位于地球和太阳之间, 但由于月球距离地球较远, 导致无法完全遮挡太阳, 而是在太阳中心形成一个环状的光环。根据图片, 太阳位于左侧, 月球位在中间, 地球位于右侧, 这表明月球处在地球和太阳之间, 这是正确的位置顺序, 以形成日环食。因此, 根据所提供的图片, 该题目描述的月球、地球和太阳的位置关系正确。答案是: A

GPT-4V (Answer) : A.

Error Category: Perception Error
Error Analysis: There is a clear positional error in GPT's understanding of the image. The moon is depicted on the outermost edge in the image, but GPT mistakenly believes the moon is positioned between the Earth and the Sun. This erroneous analysis of relative positions leads to incorrect answers.

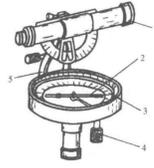
Ground Truth: B.

Figure 10. A sample perceptual error case (subject: geography, language: Chinese).

Architecture: Mapping **ENGLISH**

Question: <image_1> The figure shows a compass, and the part labeled as 2 is called what?

Options:
A. Telescope
B. Support
A. Clamp screw
B. Magnetic needle



<image_1>

GPT-4V: To answer the question, we need to identify each labeled part of the compass based on its visual characteristics and typical functions in a compass:
1. Label 1 is pointing to a long cylindrical part, which appears to be a sighting device used to align the compass with a specific object or direction. This is typically known as the telescope or sighting scope.
2. Label 2 is pointing to the arm that supports the sighting device over the body of the compass. This is consistent with the general function of a support structure, which is to hold another part in position.
3. Label 3 is pointing to the main body of the compass, which seems to encase the compass's directional mechanisms and the glass through which one observes the magnetic needle.
4. Label 4 indicates a part at the bottom of the compass, which typically refers to the mechanism that allows the compass to be leveled or adjusted for stability, often involving screws or feet.
5. Label 5 points to the circular scale around the outside of the compass, likely used to measure angles or bearings.
Now, let's evaluate each option with the above analysis:
A. Telescope - While label 1 would be associated with the telescope, this is not the part marked as 2.
B. Support - Considering the function of the part labeled 2, this option makes the most sense as it is the structure supporting the telescope.
C. Clamp screw - This is typically a component used to secure something tightly in place, and the part labeled 2 does not seem to have a screw element visible, so this isn't the most appropriate choice.
D. Magnetic needle - Label 3 points to the body of the compass where the magnetic needle would be located, not label 2.
Based on the analysis, the most appropriate term for the part labeled as 2 would be the support. The correct answer is: B

GPT-4V (Answer) : B.

Error Category: Perceptual Error
Error Analysis: A typical identical and location error. GPT identified the position labeled 2 as the support above the compass, but it actually points to the compass needle.

Ground Truth: D.

Figure 11. A sample perceptual error case (subject: architecture, language: English).

Math: Algebra

CHINESE

Question: 设行列式<image_1>, 则D₁的值为?

Options:

A. -15
B. -6
C. 6
D. 15

$$D = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = 3, D_1 = \begin{vmatrix} a_{11} & 5a_{11} + 2a_{12} & a_{13} \\ a_{21} & 5a_{21} + 2a_{22} & a_{23} \\ a_{31} & 5a_{31} + 2a_{32} & a_{33} \end{vmatrix}$$

<image_1>

GPT-4V: 为了计算行列式D₁的值, 我们可以利用行列式的性质。根据题目中已知的信息, D的值为-3。现在我们需要计算D₁的值。D₁可以看作是通过将D的第二行的每个元素乘以5后再加上第一行的相应元素的两倍来得到的一个新行列式。这个变换可以分成两步: 1. 将原行列式D的第二行的每个元素乘以5。2. 将原行列式D的第一行的每个元素乘以2后加到新的第二行对应的元素上。由于行列式的一个性质是当你将一个行(或列)乘以一个常数时, 行列式的值也会乘以这个常数, 所以在第一步, 将第二行每个元素乘以5后, D的值会乘以5。因此, 步骤1的行列式的值为 5 × (-3) = -15。然而, 在步骤2, 我们只是将第一行的两倍加到第二行上, 并没有改变行列式的值。在行列式中, 你可以将一行的若干倍加到另一行而不改变行列式的值。因此, 步骤2不会影响行列式的值。结合这两步, 行列式D₁的值就是第一步的结果, 即-15。所以答案是: A. -15

GPT-4V (Answer) : A.

Error Category: Perception Error
Error Analysis: GPT recognized '3' in an image as '-3', which is a typical OCR error.

Ground Truth: C.

Figure 12. A sample perceptual error case (subject: math, language: Chinese).

Math: Algebra

ENGLISH

Question: Given the determinant <image_1>, the value of D₁ is ?

Options:

A. -15
B. -6
C. 6
D. 15

$$D = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = 3, D_1 = \begin{vmatrix} a_{11} & 5a_{11} + 2a_{12} & a_{13} \\ a_{21} & 5a_{21} + 2a_{22} & a_{23} \\ a_{31} & 5a_{31} + 2a_{32} & a_{33} \end{vmatrix}$$

<image_1>

GPT-4V: To find the value of D₁, we need to compare the determinants D and D₁. In the given image, there is a determinant D and a modified determinant D₁. The determinant D is a 3x3. The determinant D₁ is the same 3x3 matrix, but with its second column transformed by multiplying each element by 5 first and then adding the first column to it. Now, recall that if we multiply a column in a determinant by a constant, the value of the determinant is multiplied by that constant. Additionally, if we add a multiple of one column to another column in a determinant, the value of the determinant remains unchanged. The second column of D₁ can be thought of as being formed by multiplying the first column of D by 5 and adding it to the second column multiplied by 2. This means that D₁ can be thought of as the original determinant D being added to a determinant where the first column is multiplied by 5. The first step of this process (multiplying the first column by 5) would increase the determinant's value 5 times, so we can think of this as: D₁ = 5 × D, addition of the first column to the second column, which doesn't change the determinant's value. Therefore, we have: D₁ = 5 × (-3) = -15. The correct answer is A. -15. Thus, the final response is: A.

GPT-4V (Answer) : A.

Error Category: Perception Error
Error Analysis: GPT recognized '3' in an image as '-3', which is a typical OCR error.

Ground Truth: C.

Figure 13. A sample perceptual error case (subject: math, language: English).

✗

Math: Algebra

GERMAN

Question: Gegeben die Determinante $\langle \text{image}_1 \rangle$, dann ist der Wert von D_1 ?

Options:

- A. -15
- B. -6
- C. 6
- D. 15

$$D = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = 3, D_1 = \begin{vmatrix} a_{11} & 5a_{11} + 2a_{12} & a_{13} \\ a_{21} & 5a_{21} + 2a_{22} & a_{23} \\ a_{31} & 5a_{31} + 2a_{32} & a_{33} \end{vmatrix}$$

$\langle \text{image}_1 \rangle$

GPT-4V: Um den Wert der Determinante D_1 zu berechnen, basierend auf der Information, dass die Determinante D den Wert -3 hat, müssen wir die Beziehung zwischen beiden verstehen. Die Determinante D ist definiert durch eine 3×3 Matrix mit den Elementen (a_{ij}) , während D_1 eine modifizierte Version dieser Matrix ist, bei der die erste Spalte durch die linearen Kombinationen $(5a_{i1} + 2a_{i2})$ ersetzt wurde (für $i = 1, 2, 3$). Zunächst müssen wir uns an eine wichtige Eigenschaft von Determinanten erinnern: Die Determinante einer Matrix ändert sich linear in Bezug auf die Zeilen und Spalten. Das heißt, wenn eine Spalte (oder Zeile) einer Matrix eine Summe von zwei Vektoren ist, ist die Determinante der gesamten Matrix die Summe der Determinanten der beiden entsprechenden Matrizen, eine mit der ersten und die andere mit der zweiten Vektor in der betrachteten Spalte. Daher können wir die Determinante D_1 aufspalten in die Summe zweier Determinanten. Die erste Determinante auf der rechten Seite ist einfach das 5-fache der Determinante (D) (da jede Zeile der ersten Spalte von D mit 5 multipliziert wird), daher der Wert $(5 \cdot (-3)) = -15$. Die zweite Determinante ist jedoch 0, da die erste und die zweite Spalte gleich sind, was bedeutet, dass die Spalten linear abhängig sind und die Determinante einer Matrix mit linear abhängigen Spalten immer 0 ist. Somit ist der Wert von $(D_1 = -15 + 0 = -15)$. Die korrekte Antwort ist also: A. -15.

GPT-4V (Answer) : A.

Error Category: Perception Error

Error Analysis: GPT recognized '3' in an image as '-3', which is a typical OCR error.

Ground Truth: C.

Figure 14. A sample perceptual error case (subject: math, language: German).

✗

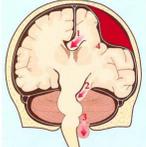
Laboratory Medicine: Pathology

CHINESE

Question: $\langle \text{image}_1 \rangle$ 如图为脑疝模式图，数字2对应于以下哪个选项:

Options:

- A. 扣带回疝
- B. 小脑扁桃体疝
- C. 海马体回疝
- D. 占位病变



$\langle \text{image}_1 \rangle$

GPT-4V: 在这幅图中，我们可以看到大脑的剖面图，标有数字1到4的不同区域。数字2位于大脑的下方区域，这块区域被认为是小脑的部分，并且从它的位置来看，它指向的是小脑的一个特定部分，这个部分就在大脑后方和脑干之间的区域。根据选项：A. 扣带回疝——这通常描述的是大脑内部的一种疝压现象，但并不在小脑区域。B. 小脑扁桃体疝——这个选项描述的是小脑扁桃体（cerebellar tonsils）通过颅底的开口（例如枕骨大孔）向下移位的情况，这通常发生在颅内压增高时。C. 海马体回疝——这描述的是大脑海马体的一个部分发生的疝压事件，跟小脑无关。D. 占位病变——这个选项通常是指大脑内部的肿瘤、血肿或其他占位过程，但它并不是特指疝的类型。因此，结合图中数字2所指的位置和选项的描述，最合适的选项是B. 小脑扁桃体疝。答案：B

GPT-4V (Answer) : B.

Error Category: Perceptual Error, Lack of Knowledge

Error Analysis: GPT mistook the hippocampus in the image for the cerebellum, which can be attributed to both visual perception and domain-specific knowledge.

Ground Truth: C.

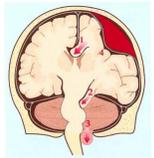
Figure 15. A sample perceptual error case (subject: laboratory medicine, language: Chinese).



Laboratory Medicine: Pathology

ENGLISH

Question: <image_1> As shown in the Brain herniation model diagram, number 2 corresponds to which of the following options:
Options:
A. Uncinate herniation
B. Cerebellar tonsillar herniation
C. Hippocampal herniation
D. Occupational disease change



<image_1>

GPT-4V: The image you provided is a diagram illustrating types of brain herniation. Let 's analyze each option in conjunction with the number 2 marked in the diagram. A. Uncinate herniation refers to the herniation of the uncus, which is part of the hippocampal gyrus in the medial temporal lobe. This type of herniation typically affects the area near the tentorium, which would involve lateral displacement. B. Cerebellar tonsillar herniation involves the downward displacement of the cerebellar tonsils through the foramen magnum. This is typically indicated at the base of the brain and is associated with the marker that would be at the lowest part of the brain in the diagram. C. Hippocampal herniation, which is similar to uncinatate herniation, would involve the medial temporal lobe structures. This event also belongs to the category of transtentorial herniation and affects areas around the tentorium, similar to option A. D. Occupational disease change is not a type of brain herniation. Instead, this refers to conditions acquired due to exposures in the workplace. It does not fit the context of the diagram. Considering the options and looking at the diagram associated with number 2, we see that there is a downward displacement happening near the base of the brain (near the cerebellum), into the opening we know as the foramen magnum. Based on the location and direction of the herniation in the diagram, it resembles cerebellar tonsillar herniation rather than the others. Therefore, the correct answer is: B. Cerebellar tonsillar herniation. B

GPT4V (Answer) : B.

Error Category: Perceptual Error, Lack of Knowledge
Error Analysis: GPT mistook the hippocampus in the image for the cerebellum, which can be attributed to both visual perception and domain-specific knowledge.

Ground Truth: C.

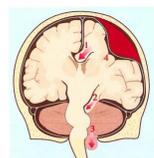
Figure 16. A sample perceptual error case (subject: laboratory medicine, language: English).



Laboratory Medicine: Pathology

GERMAN

Question: <image_1> Das Bild zeigt das Muster des Hirnvorfalles die Nummer 2 der folgenden Option:
Options:
A. Falx cerebri Herniation
B. Kleinhirntonsillen Herniation
C. Hippocampus Herniation
D. Raumfordernde Läsion



<image_1>

GPT-4V: Um die vorliegende Frage zu beantworten, werde ich jede der Optionen analysieren und sie mit der Nummer 2 aus dem Bild vergleichen, die einen bestimmten Typ einer Hirnherniation anzeigen soll. Option A: "Falx cerebri Herniation" bezieht sich auf eine Verlagerung von Gehirnteilen unter der Hirnsichel (Falx cerebri). Dies tritt meist zwischen den beiden Hemisphären auf. Auf dem Bild markiert die Nummer 2 den Bereich nahe der Basis des Gehirns, was nicht typisch für eine Falx cerebri Herniation ist. Option B: "Kleinhirntonsillen Herniation" bezeichnet eine Einengung der Kleinhirntonsillen in das Foramen magnum, welche die Schädelhöhle mit dem Spinalkanal verbindet. Auf dem Bild deutet die Position der Nummer 2, die sich am unteren Rand des Gehirns befindet und in den Wirbelkanal verlagert zu sein scheint, auf eine Kleinhirntonsillen Herniation hin. Option C: "Hippocampus Herniation" ist eine Herniation, die den Hippocampus betrifft, einem Teil des limbischen Systems tief im Gehirn. Die Position der Nummer 2 im Bild liegt nicht in der Nähe des Hippocampus, sodass diese Option unwahrscheinlich erscheint. Option D: "Raumfordernde Läsion" ist ein allgemeiner Begriff für eine Massenverlagerung im Gehirn, die sich aus vielen möglichen Ursachen ergeben kann, einschließlich Tumoren, Abszessen, Blutungen oder Entzündungen. Obwohl die Nummer 2 im Bild eine Art Verlagerung zeigt, ist die Bezeichnung "Raumfordernde Läsion" ohne zusätzliche Information zu unspezifisch, um direkt mit der Nummer 2 korreliert zu werden. Auf Basis der Analyse und der Position der Nummer 2 im Bild ist die richtige Antwort:\n\nB. Kleinhirntonsillen Herniation. Daher lautet die Antwort in Großbuchstaben: B

GPT-4V (Answer) : B.

Error Category: Perceptual Error, Lack of Knowledge
Error Analysis: GPT mistook the hippocampus in the image for the cerebellum, which can be attributed to both visual perception and domain-specific knowledge.

Ground Truth: C.

Figure 17. A sample perceptual error case (subject: laboratory medicine, language: German).

✘

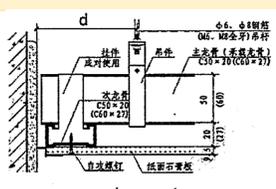
Architecture : Structural Engineering

GERMAN

Question: <image_1> Wie in der Abbildung gezeigt, was sollte der maximale Abstand d vom Hosenträger zum Ende des primären Skelettrahmens gemäß dem Standard sein?"

Options:

A. 600
B. 400
C. 300
D. 150



<image_1>

GPT-4V: Um die Frage zu beantworten, muss ich mir das Bild ansehen und nach Hinweisen oder Markierungen suchen, die auf den Standardabstand hinweisen könnten, der für den maximalen Abstand "d" vom Hosenträger zum Ende des primären Skelettrahmens erforderlich ist. Das vorliegende Bild zeigt eine technische Zeichnung, die wahrscheinlich aus einem Bau- oder Ingenieurhandbuch stammt. Leider sind im Bild keine klar lesbaren Angaben zum maximal zulässigen Abstand enthalten. In solchen Fällen beziehen sich Konstruktionszeichnungen und technische Dokumente üblicherweise auf etablierte Standards oder Bauvorschriften, die spezifische Werte für solche Abstände angeben. Da ich nicht auf externe Datenquellen oder frühere Kenntnisse der spezifischen Bauvorschriften zugreifen kann, die für diese Zeichnung relevant wären, kann ich keine definitive Antwort auf der Grundlage des Bildes allein geben. Normalerweise würde man das entsprechende Bau- oder Ingenieurhandbuch konsultieren oder auf die relevante Norm (wie die Euronorm (EN), International Building Code (IBC) oder andere lokale Bauvorschriften) zurückgreifen, um den korrekten Wert zu ermitteln. Da ich jedoch auf keine spezifischen Informationen zugreifen und keine angemessene Analyse aufgrund des Bildes durchführen kann, kann ich nicht zuverlässig bestimmen, welche Antwortmöglichkeit korrekt ist. Die Antwortmöglichkeiten sind: A. 600 B. 400 C. 300 D. 150. Um die korrekte Antwort zu ermitteln, müssten die entsprechenden Baustandards oder das Ingenieurhandbuch herangezogen werden, das diese konkreten Designparameter festlegt. Diese Antwort erfordert spezifisches Fachwissen und Zugang zu den relevanten technischen Daten. Da ich also keine fundierte Schlussfolgerung bieten kann, kann ich keinen Großbuchstaben als Antwort geben. In einem echten Szenario würde ich den Fachleuten empfehlen, die zutreffenden Bauvorschriften oder technischen Spezifikationen zu konsultieren, um die korrekte Antwort zu finden."

GPT-4V (Answer) : A.

Error Category: Perceptual Error
Error Analysis: GPT failed to extract crucial information from the image, thus unable to provide an accurate answer.

Ground Truth: C.

Figure 20. A sample perceptual error case (subject: architecture, language: German).

✘

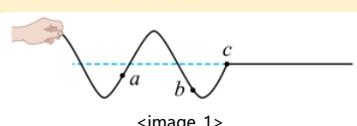
Physics: Classical Mechanics

CHINESE

Question: 取一条较长的软绳,用手握住一端拉平后连续周期性地向上、向下抖动,可以看到在绳上产生一列波, a, b, c 为绳上的质点,某时刻波刚好传播到质点 c,绳上形成的波形如图<image_1>所示,下列说法正确的是?"

Options:

A. 手开始抖动时运动方向向上
B. 之后质点 a 比质点 b 先回到平衡位置
C. 该时刻质点 a 的速度比质点 b 的速度大
D. 图中为波源开始振动后 1.5 周期时刻的波形



<image_1>

GPT-4V: 为了分析这个问题,我们可以根据图中显示的波形来考虑每个选项: A. 手开始抖动时运动方向向上: 从波的形状来判断,如果手的抖动是向上开始的,我们应该在绳的左端(即波源处)看到向上的波峰。但是,图中显示了一个向下的波谷。因此,根据这个图像,我们不能确定手是向上还是向下抖动来开始产生波形。该选项无法仅通过图中的信息来验证。 B. 之后质点 a 比质点 b 先回到平衡位置: 在波动中,质点先达到峰或谷的部位比后达到的部位先回到平衡位置。图中显示质点 a 在质点 b 的左侧,也就意味着波传播到质点 a 在质点 b 之前。因此,质点 a 将比质点 b 先回到平衡位置。这个说法是正确的。 C. 该时刻质点 a 的速度比质点 b 的速度大: 在同一个波中,波峰或波谷附近的质点移动速度是最大的,这是因为这些位置的质点具有最大的位移。质点 a 位于波谷,而质点 b 在接近平衡位置的地方,所以质点 a 的速度比质点 b 大。这个说法是正确的。 D. 图中为波源开始振动后 1.5 周期时刻的波形: 要判断周期需要知道完整一个波形的样子。以波源处的位置为基准,我们可以看到图中有一个完整的波谷和波峰,以及另一个开始形成的波谷。因此,这表示波源已经完成了完整的上下周期,并开始了第二个周期。但是,我们不能从这张图片中准确判断波源振动了 1.5 个周期。我们需要更多信息来确定周期数。这个说法无法仅通过图中的信息来验证。从以上分析来看,选项 B 和选项 C 是正确的。不过,如果题目要求选择一个最合适的答案,那么我们将选择质点 a 和质点 b 之间的对比以确定答案。这个问题最终的答案取决于题目的要求,因为有两个陈述可能是正确的,不过如果要求选择一个最佳答案,那么: B. 之后质点 a 比质点 b 先回到平衡位置。这是一个既符合波动学原理,又可以直接从波形图中验证的陈述。最终答案: B

GPT-4V (Answer) : B.

Error Category: Lack of Knowledge
Error Analysis: The knowledge that in wave motion, the point reaching the crest or trough first returns to the equilibrium position before the point reaching later applies only under the condition of a single cycle.

Ground Truth: C.

Figure 21. A sample lack of knowledge case (subject: physics, language: Chinese).

Energy : Thermodynamics

CHINESE

Question: 其制冰过程可简化为图中的循环过程，其中横轴为温度 T ，纵轴为压强 p ；过程A→B：一定量的二氧化碳在压缩机的作用下变为高温高压的超临界态（一种介于液态和气态之间，分子间有强烈相互作用的特殊状态）；过程B→C：二氧化碳在冷凝器中经历一恒压过程向外放热而变成高压液体；过程C→D：二氧化碳进入蒸发器中蒸发，进而使与蒸发器接触的水降温凝固；过程D→A：二氧化碳经历一恒压过程回到初始状态。下列说法正确的是哪项？

Options:

A. 过程A→B中，每个二氧化碳分子的动能都将增大
 B. 过程B→C中，二氧化碳始终遵循理想气体的实验定律
 C. 过程D→A中，若二氧化碳可看作理想气体，则该过程中二氧化碳将吸热
 D. 整个循环过程中，热量从低温水向高温二氧化碳传递，违反热力学第二定律

<image_1>

GPT-4V: 在研究这道题目时，我们会按照题干所描述的过程A→B、B→C、C→D和D→A逐步分析，并对对应给出的四个选项A、B、C和D。过程A→B是一个压缩过程，二氧化碳被压缩成高温高压的超临界态。压缩过程中，分子的动能会增加，因此选项A表述“过程A→B中，每个二氧化碳分子的动能都将增大”是正确的。选项B说“过程B→C中，二氧化碳始终遵循理想气体的实验定律”，但这个过程中二氧化碳是超临界态，并且在冷凝过程中会由气态变为液态，所以不会始终遵循理想气体定律。因此选项B是错误的。选项C说“过程D→A中，若二氧化碳可看作理想气体，则该过程中二氧化碳将吸热”。过程D→A是一个膨胀过程，在这一过程中，二氧化碳从液态变为气态，即蒸发过程。理想气体在膨胀的时候，如果外界不做功，是需要吸热的。因此，假设二氧化碳可以看作理想气体，选项C的表述是正确的。选项D说“整个循环过程中，热量从低温水向高温二氧化碳传递，违反热力学第二定律”。这个说法是错误的，因为正确的循环制冷过程是要从低温物体接收热量，并向高温处排放热量，与热力学第二定律相符。热力学第二定律描述了热能传递的可逆性方向，热量自然流向是从高温向低温。综上所述，正确的答案是A（选项A的表述）。答案：A

GPT-4V (Answer) : A.

Error Category: Lack of Knowledge
Error Analysis: GPT's oversight of the fact that an increase in temperature leads to an increase in molecular kinetic energy refers to the average kinetic energy of molecules. This is a factual error resulting from the neglect of objective principles.

Ground Truth: C.

Figure 22. A sample lack of knowledge case (subject: energy, language: Chinese).

Electronics: Electronics Circuits

ENGLISH

Question: <image_1> In the circuit shown, what is the capacitor voltage u_c at $t=0_+$ when the switch is open?

Options: A. 2V B. 3V C. 4V D. 0V

Ground Truth: B.

<image_1>

GPT-4V: To determine the capacitor voltage u_c right after the switch is opened at $(t = 0)$, we need to consider the state of the circuit immediately before $(t = 0)$, since capacitors cannot change their voltage instantaneously. Here's the step-by-step analysis: 1. Identify the state of the capacitor before the switch opens: Since the capacitor is connected across a 4V battery and the circuit has been stable before $(t = 0)$, the capacitor is fully charged to the voltage of the battery. 2. Recognize that the voltage across a capacitor cannot change suddenly because it would require infinite current, which is not physically possible. Therefore, the voltage across the capacitor just after the switch is opened u_c will be the same as the voltage across it just before the switch is opened. 3. Determine the voltage across the capacitor before $(t = 0)$: Since the capacitor was connected across a 4V battery before $(t = 0)$, it will be charged to 4V. Therefore, the answer is: C.

GPT-4V (Answer) : C.

Error Category: Reasoning Error
Error Analysis: GPT only considers the power supply on the left and does not consider the power supply on the right. In the process of analysis, critical conditions that have been obtained are forgotten.

Figure 23. A typical case of reasoning error (subject: electronics, language: English).

⊗

Physics: Classical Mechanics

ENGLISH

Question: Take a relatively long flexible rope, hold one end and pull it tight, then continuously shake it up and down, a series of waves can be seen on the rope, with \$a, b, c\$ being mass points on the rope. At a certain moment, the wave just reaches mass point \$c\$, and the wave shape formed on the rope is as shown in <image_1>, the correct statement(s) is(are)

Options:

A. The hand starts shaking in the upward direction
 B. Afterward, mass point \$a\$ returns to the equilibrium position before mass point \$b\$
 C. At that moment, the speed of mass point \$a\$ is greater than that of mass point \$b\$
 D. The wave shape in the figure is at 1.5 periods after the wave source starts oscillating

<image_1>

GPT-4V: Let's analyze the options provided, based on the wave shape shown in the image. A. The hand starts shaking in the upward direction. Since the wave starts with an upward motion close to the hand (left side of the image), we can infer that the hand also started by shaking in the upward direction. This is because the first part of the wave will reflect the initial motion of the hand. This makes option A correct. B. Afterward, mass point \$a\$ returns to the equilibrium position before mass point \$b\$. The wave travels from left to right, and since mass point \$a\$ is to the left of mass point \$b\$, it oscillates first. After the wave passes, \$a\$ will return to the equilibrium position before \$b\$ does because the wave will reach and move \$a\$ before it reaches \$b\$. This makes option B correct. C. At that moment, the speed of mass point \$a\$ is greater than that of mass point \$b\$. Speed in wave motion is related to the slope of the wave at that point. The steeper the slope of the wave at a point, the faster the mass point will be moving through its equilibrium position. At the given instant, mass point \$a\$ is at a steeper part of the wave than mass point \$b\$, suggesting that it is moving faster. Hence, option C is correct. D. The wave shape in the figure is at 1.5 periods after the wave source starts oscillating. Analyzing the wave shown in the image, there is one complete wave cycle (a crest and a trough), plus a half cycle (either a crest or a trough). This constitutes 1.5 periods. Therefore, option D is also correct. All options A, B, C, and D seem to be correct based on the image provided. The answer, therefore, is A, B, C, and D. However, you have not provided any information about the initial phase with which the person starts to shake the rope (up or down) or the convention as to what the starting direction should be considered. So, based on conventional assumptions, my answer is: ABCD.

GPT-4V (Answer) : B.

Error Category: Reasoning Error
Error Analysis: Although GPT correctly extracted the information from the image and possesses basic knowledge to infer the direction of wave propagation, it failed to infer the direction of wave propagation and treated it as unknown information in its inference.

Ground Truth: C.

Figure 24. A sample reasoning error case (subject: physics, language: English).

⊗

Geography: Geophysics

ENGLISH

Question: Judge true or false, during a solar eclipse, the correct positional relationship between the Moon, Earth, and Sun is <image_1>:

Options:

A. True
 B. False

<image_1>

GPT-4V: To analyze the question, we first need to understand the phenomenon of a solar eclipse. A solar eclipse occurs when the Moon passes between the Earth and the Sun, blocking all or part of the Sun's light from reaching the Earth. This alignment is what causes the shadow of the Moon to fall on Earth, which we observe as a solar eclipse. Now, let's look at the image provided. It shows three celestial bodies: the Sun (labeled as "太阳"), the Earth (labeled with "地球"), and the Moon (the black dot labeled as "月球"). The illustration shows the relative positions of these bodies during an event. The Sun is on the left, shining its light toward the Earth. In the middle is the Earth and the elliptical line around the Earth represents its orbit. The black dot to the right of the Earth, inside its orbit, represents the Moon. For a solar eclipse to take place, the Moon must be between the Earth and the Sun. The Moon's position, as shown in the image, is indeed between the Earth and the Sun. Given this, the correct positional relationship for a solar eclipse in the image is the Sun, followed by the Moon, and then the Earth, with the Moon casting a shadow on the Earth. Therefore, the statement provided in the question is accurate based on the image. The answer to the question is: A. True. So the capital letter you asked for at the end would be: A.

GPT-4V (Answer) : A.

Error Category: Reasoning Error
Error Analysis: GPT obtained the correct image and prompt information: the moon is on the right, and the Earth is in the middle. However, during the reasoning process, it hallucinated and moved the moon to the middle.

Ground Truth: B.

Figure 25. A sample reasoning error case (subject: geography, language: English).

✖
Geography: Geophysics
GERMAN

Question: Richtig oder falsch, während einer Sonnenfinsternis ist die Position zwischen Mond, Erde und Sonne korrekt in <image_1>:

Options:

A. Richtig

B. Falsch



<image_1>

GPT-4V: Um diese Frage zu beantworten, analysieren wir zunächst das Bild. Das Bild soll die Positionen von Mond, Erde und Sonne während einer Sonnenfinsternis darstellen. Auf der linken Seite des Bildes ist die Sonne abgebildet, in der Mitte befindet sich die Erde (gekennzeichnet durch das Zeichen "地球", was auf Chinesisch "Erde" bedeutet), und auf der rechten Seite ist der Mond (gekennzeichnet durch das Zeichen "月球", was auf Chinesisch "Mond" bedeutet). Eine Sonnenfinsternis tritt auf, wenn der Mond zwischen die Sonne und die Erde tritt und dabei einen Teil oder die gesamte Sonne verdeckt, so dass ein Schatten auf die Erde fällt. Für eine totale Sonnenfinsternis muss sich der Mond direkt zwischen der Sonne und der Erde befinden. Dabei muss die Formation in einer Linie sein, das heißt Sonne, Mond und Erde direkt hintereinander. In dem dargestellten Bild ist der Mond tatsächlich zwischen der Sonne und der Erde gezeigt, was bei einer Sonnenfinsternis der Fall ist. Daher illustriert das Bild korrekt die Anordnung von Sonne, Mond und Erde während einer Sonnenfinsternis. Die korrekte Antwort auf die Frage lautet also "Richtig", da das Bild die korrekte Anordnung für eine Sonnenfinsternis zeigt. Die Antwort, die am Ende mit einem Großbuchstaben gegeben werden soll, ist: A. Richtig

GPT-4V (Answer) : A.

Error Category: Reasoning Error

Error Analysis: GPT obtained the correct image and prompt information: the moon is on the right, and the Earth is in the middle. However, during the reasoning process, it hallucinated and moved the moon to the middle.

Ground Truth: B.

Figure 26. A sample reasoning error case (subject: geography, language: German).