MathScape: Benchmarking Multimodal Large Language Models in Real-World Mathematical Contexts

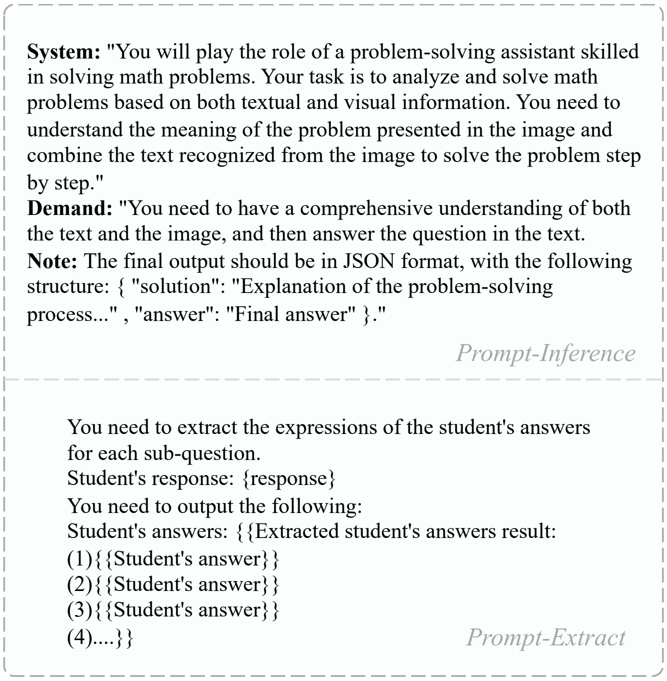
Supplementary Material

Hao Liang[♠], Linzhuang Sun[♣], Minxuan Zhou[♡], Meiyi Qiang[♠], Mingan Lin[♡], Tianpeng Li[♡], Fan Yang[♡], Zenan Zhou[♡], Wentao Zhang[♠]

[♠]Peking University [♡]Baichuan Inc. [♣]University of Chinese Academy of Sciences

1 Prompt for inference, extracting and Scoring Answers

We summarize the prompt for scoring answers in Figure 2.



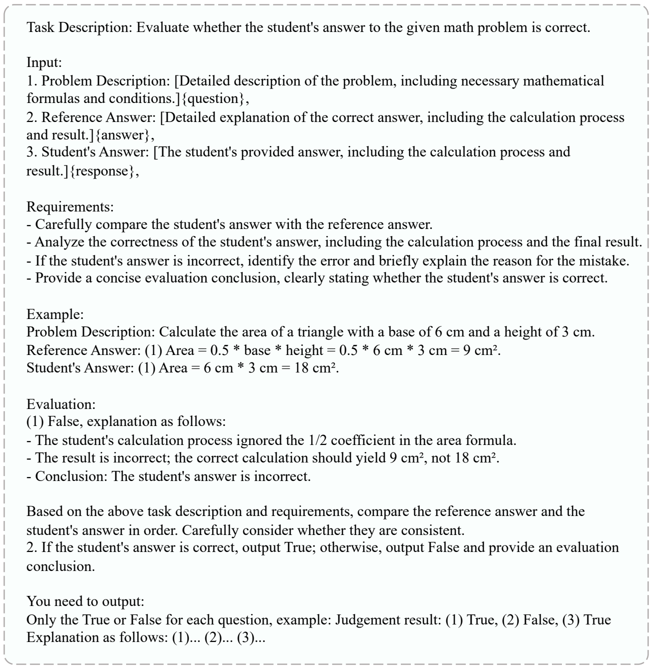
System: "You will play the role of a problem-solving assistant skilled in solving math problems. Your task is to analyze and solve math problems based on both textual and visual information. You need to understand the meaning of the problem presented in the image and combine the text recognized from the image to solve the problem step by step."
Demand: "You need to have a comprehensive understanding of both the text and the image, and then answer the question in the text.
Note: The final output should be in JSON format, with the following structure: { "solution": "Explanation of the problem-solving process..." , "answer": "Final answer" }."

Prompt-Inference

You need to extract the expressions of the student's answers for each sub-question.
Student's response: {response}
You need to output the following:
Student's answers: {{Extracted student's answers result:
(1){{Student's answer}}
(2){{Student's answer}}
(3){{Student's answer}}
(4)....}}

Prompt-Extract

Figure 1: Prompts for inference and extracting answers.



Task Description: Evaluate whether the student's answer to the given math problem is correct.

Input:
1. Problem Description: [Detailed description of the problem, including necessary mathematical formulas and conditions.]{question},
2. Reference Answer: [Detailed explanation of the correct answer, including the calculation process and result.]{answer},
3. Student's Answer: [The student's provided answer, including the calculation process and result.]{response},

Requirements:
- Carefully compare the student's answer with the reference answer.
- Analyze the correctness of the student's answer, including the calculation process and the final result.
- If the student's answer is incorrect, identify the error and briefly explain the reason for the mistake.
- Provide a concise evaluation conclusion, clearly stating whether the student's answer is correct.

Example:
Problem Description: Calculate the area of a triangle with a base of 6 cm and a height of 3 cm.
Reference Answer: (1) Area = 0.5 * base * height = 0.5 * 6 cm * 3 cm = 9 cm^2.
Student's Answer: (1) Area = 6 cm * 3 cm = 18 cm^2.

Evaluation:
(1) False, explanation as follows:
- The student's calculation process ignored the 1/2 coefficient in the area formula.
- The result is incorrect; the correct calculation should yield 9 cm^2, not 18 cm^2.
- Conclusion: The student's answer is incorrect.

Based on the above task description and requirements, compare the reference answer and the student's answer in order. Carefully consider whether they are consistent.
2. If the student's answer is correct, output True; otherwise, output False and provide an evaluation conclusion.

You need to output:
Only the True or False for each question, example: Judgement result: (1) True, (2) False, (3) True
Explanation as follows: (1)... (2)... (3)...

Figure 2: Prompt used for scoring answers.

2 Case Study & Key Challenges

Our MathScape benchmark introduces several challenges for multiple models. In this section, we explore the primary reasons models provide incorrect answers to image-text mathematical problems. These errors primarily arise from difficulties in understanding and interpreting the provided information. We categorize these challenges as follows:

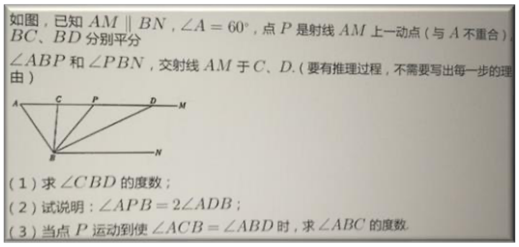
Incorrect Information from the Image: One of the most frequent errors arises when models fail to correctly extract relevant information from the image. For instance, as demonstrated in Case Study 1 in Figure 3, the image clearly shows that $\angle PBD \neq \angle ABD$; however, the model generates an incorrect proof due to inaccurate information extracted from the image.

Incorrect Spatial Positioning: This challenge pertains to the model's inability to accurately interpret spatial relationships. As illustrated in Case Study 2 in Figure 3, even GPT-4V fails to recognize that the origin lies to the left of point 1, leading to an incorrect conclusion.

Limited Reasoning Capacity: This challenge stems from the insufficient logical reasoning capabilities of LLMs. Even when the image data is correctly extracted, the LLM may still produce erroneous conclusions due to its limited reasoning capacity. As shown in Case Study 3 in Figure 3, the LLM fails to solve a complex problem correctly and makes logical and computational mistakes during the reasoning process.

Overall, LLMs face significant challenges in mathematical reasoning, compounded by the need to process all visual information. This limitation affects their problem-solving abilities, especially when interpreting complex patterns such as geometric configurations or charts. To improve performance, future research should focus on developing models with stronger mathematical reasoning skills and enhanced pattern recognition capabilities.

(a)



如图，已知 $AM \parallel BN$，$\angle A=60^\circ$，点 P 是射线 AM 上一动点（与 A 不重合），BC、BD 分别平分 $\angle ABP$ 和 $\angle PBN$，交射线 AM 于 C、D.（要有推理过程，不需要写出每一步的理由）

（1）求 $\angle CBD$ 的度数；

（2）试说明：$\angle APB=2\angle ADB$；

（3）当点 P 运动到使 $\angle ACB=\angle ABD$ 时，求 $\angle ABC$ 的度数.

Given the figure, it is known that AM∥BN, ∠A=60∘, point P is a moving point on the ray AM (not coincident with A), BC and BD bisect ∠ABP and ∠PBN respectively. The ray AM intersects C and D. (There needs to be a reasoning process, but it is not necessary to write out every step.)

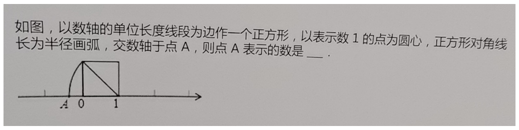
1. Find the measure of ∠CBD
2. Prove that ∠APB=2∠ADB
3. When point P moves such that ∠ACB=∠ABD, find the measure of ∠ABC.

(b) First, since BD is the bisector of ∠PBN, similarly, we can conclude that ∠PBD=∠ABD=60°. Then, because ∠APB=∠ABP+∠ABD, we know that ∠ABP=∠ABC (since BC bisects ∠ABP), which equals 60°. Therefore, ∠APB=60°+60°=120°. Now, we can calculate ∠ADB. Since ∠ABD=60°, and we also know that ∠BAD=60°, in the isosceles triangle ΔABD, ∠ADB=180°−∠BAD−∠ABD=180°−60°−60°=60∘ Therefore, we can say ∠APB=120°=2×60°=2∠ADB

(c) **Judge:** The proof is complete incorrect. **Difficulty:** Recognition of spatial angle relationships, spatial deduction and proof.

Case Study 1

(a)



如图，以数轴的单位长度线段为边作一个正方形，以表示数 1 的点为圆心，正方形对角线长为半径画弧，交数轴于点 A，则点 A 表示的数是 ___ .

When actually measuring the height of a mountain, you can measure the relative height between adjacent observable points among several observation points and then use these relative heights to calculate the mountain's height. The table below records some of the measurement data (e.g., A-C=90 shows A is 90 meters higher than C):

(b) First, the problem describes a square with a side length equal to the unit length representing the number 1 on the number line. Let's denote the side length of this square as 1 unit.

Next, the problem states that the point representing the number 1 is the center of a circle, and the radius of the arc is equal to the length of the diagonal of the square. We know that the length of the diagonal of a square is$\sqrt{2}$ times the side length, so in this case, the diagonal length is $\sqrt{2}$.

Therefore, with 1 as the center and $\sqrt{2}$ as the radius, the length of this arc represents the distance on the number line. Simply put, the arc intersects the number line at point A, and the distance from point A to the origin (0) on the number line is $\sqrt{2}$.

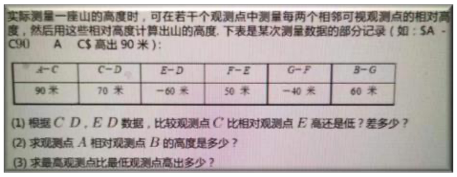
So, the number represented by point A is 1 plus $\sqrt{2}$, which means point A represents the number 1+ $\sqrt{2}$.

Final Answer: 1+ $\sqrt{2}$

(c) **Judge:** The answer is incorrect. **Difficulty:** Weak spatial ability and spatial position reasoning skills.

Case Study 2

(a)



实际测量一座山的高度时，可在若干个观测点中测量每两个相邻可视观测点的相对高度，然后用这些相对高度计算出山的高度. 下表是某次测量数据的部分记录（如：$A - C90 A C$ 高出 90 米）：

A−C	C−D	E−D	F−E	G−F	B−G
90 米	70 米	−60 米	50 米	−40 米	60 米

(1) 根据 C D，E D 数据，比较观测点 C 比相对观测点 E 高还是低？差多少？

(2) 求观测点 A 相对观测点 B 的高度是多少？

(3) 求最高观测点比最低观测点高出多少？

As shown in the figure, using a unit length segment on the number line as the side of a square, with the point representing the number 1 as the center of a circle, a semicircle is drawn with a radius equal to the side length of the square. The semicircle intersects the number line at point A. The number represented by point A is ____.

1) Based on the data of C-D and E-D, is observation point C higher or lower than observation point E? By how much?
2) How much higher is observation point A compared to observation point B?
3) How much higher is the highest observation point compared to the lowest observation point?

(b)

Based on the information provided in the image, let's first interpret the data in the table:
A − C = 90 meters, C − D = 70 meters, E − D = -60 meters, F − E = 50 meters, G − F = -40 meters, B − G = 60 meters
Now, let's solve the questions step by step:

(1) First, let's compare the heights of observation points C and E. Since E − D is -60 meters, this means point D is 60 meters higher than point E. C − D is 70 meters, meaning point C is 70 meters higher than point D. Therefore, the height of point C relative to point E is the 70 meters that C is higher than D plus the 60 meters that D is higher than E, **for a total of 130 meters.** Thus, observation point C is 130 meters higher than observation point E.

(2) We can calculate the height of point A relative to point B by combining the consecutive height differences:
A − C = 90 meters
C − B = (C − D) + (D − E) + (E − F) + (F − G) + (B − G) = 70 + 60 + 50 + 40 + 60 = 280 meters
So, the height of point A relative to point B is A − C + C − B = 90 + **280 = 370 meters**.

(3) To find the highest and lowest observation points, we can observe the height differences relative to each other. From the data provided, the highest observation point is point A, and the lowest observation point is point E (as no point is lower than E). From the previous calculations, we already know that point A is 130 meters higher than point E.
Since point A is 90 meters higher than point C, and point C is 130 meters higher than point E, we have:
The highest point A is 220 meters higher than the lowest point E:
A − E = A − C + C − E = 90 + 130 = 220 meters.
This concludes the solution process. Here are the answers to each question:
(1) The height difference by which observation point C is higher than observation point E: **130 meters**
(2) The height of observation point A relative to observation point B: **370 meters**
(3) The height difference between the highest and the lowest observation points: **220 meters**

(c) **Judge:** The answer is partial incorrect. **Difficulty:** The reasoning chain is too long; involves six steps of reasoning; Requires accurate memory ability.

Case Study 3

Figure 3: Illustrations of three representative case studies. Each example includes the original question (a), the response from GPT-4V (b), and our final judgment (c). Red markers highlight the erroneous portions of the model's answer.