MM-MATH: Advancing Multimodal Math Evaluation with Process Evaluation and Fine-grained Classification

Kai Sun[*], **Yushi Bai**[*], **Ji Qi, Lei Hou**[†], **Juanzi Li**
Tsinghua University, Beijing, 100084, China
{sunk20, bys22}@mails.tsinghua.edu.cn

Abstract

To advance the evaluation of multimodal math reasoning in large multimodal models (LMMs), this paper introduces a novel benchmark, MM-MATH. MM-MATH consists of 5,929 open-ended middle school math problems with visual contexts, with fine-grained classification across difficulty, grade level, and knowledge points. Unlike existing benchmarks relying on binary answer comparison, MM-MATH incorporates both outcome and process evaluations. Process evaluation employs LMM-as-a-judge to automatically analyze solution steps, identifying and categorizing errors into specific error types. Extensive evaluation of ten models on MM-MATH reveals significant challenges for existing LMMs, highlighting their limited utilization of visual information and struggles with higher-difficulty problems. The best-performing model achieves only 31% accuracy on MM-MATH, compared to 82% for humans. This highlights the challenging nature of our benchmark for existing models and the significant gap between the multimodal reasoning capabilities of current models and humans. Our process evaluation reveals that diagram misinterpretation is the most common error, accounting for more than half of the total error cases, underscoring the need for improved image comprehension in multimodal reasoning. The code and dataset are available at `https://github.com/kge-sun/MM-Math`.

1 Introduction

Due to their exceptional performance in handling complex text and images, large multimodal models (LMMs) such as GPT-4V (OpenAI, 2023) and Claude-3 (Anthropic, 2024) have garnered significant interest in both industry and academia. Previous studies suggest that they still underperform on multimodal math reasoning tasks (Chen et al., 2021; Lu et al., 2023; Zhang et al., 2024), as such tasks require understanding multimodal information and interleaving reasoning within this information (Lightman et al., 2023). To further advance LMM's mathematical capabilities, we believe the following two issues urgently need addressing: **(1) What are the specific reasons that lead to the model's mistakes**, such as misunderstanding the diagram or errors in reasoning? **(2) How does the model perform across different categories of multimodal math problems**, and which specific types of problems does the model excel at or struggle with?

In this paper, we introduce **MM-MATH** benchmark to provide a more fine-grained and reliable assessment of LMMs' multimodal math capability. MM-MATH comprises a total of 5,793 open-ended multimodal math problems from middle school. We show an overview of the design of MM-MATH in Figure 1. To address the aforementioned issue (1), MM-MATH combines traditional outcome evaluation (comparing the model's answer to groundtruth and reaching binary result) with **process evaluation**. Process evaluation involves using LMM-as-a-judge (Zheng et al., 2023) to automatically identify errors in the model's output process and categorize the causes of these errors. Concretely, we employ GPT-4V (OpenAI, 2023) to compare the step-by-step solution generated by the model with our annotated groundtruth solution, and identify the first error in the model's process to determine the main reason that leads to a wrong answer. We categorize the causes of LMM's errors into four types in Figure 1, including *diagram misinterpretation*, *reasoning error*, *calculation error*, and *textual condition misunderstanding*.

In response to issue (2), MM-MATH includes **fine-grained classification**, where the problems are classified along three dimensions: difficulty, grade level, and knowledge points, to evaluate the breadth, depth, and specific knowledge for math
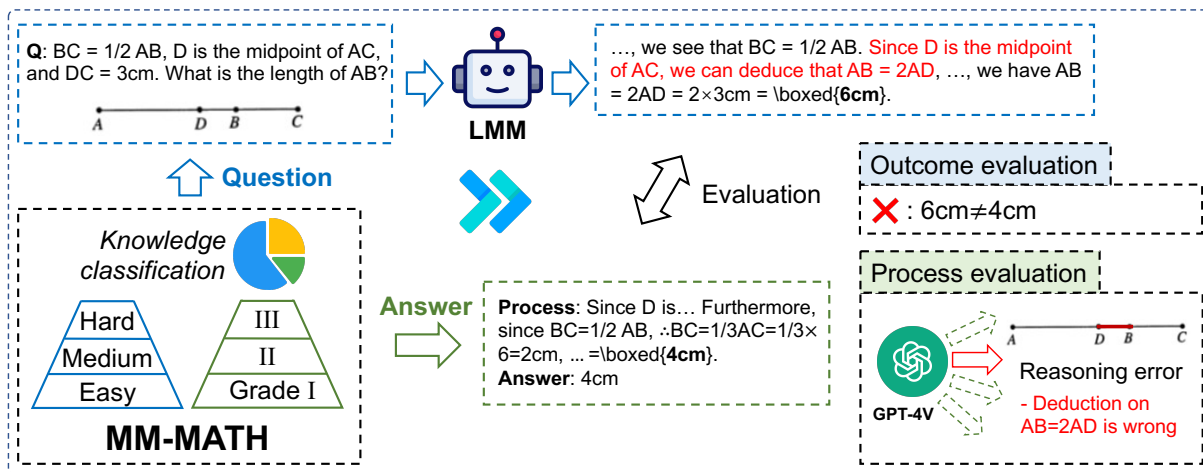
---

[*]Equal contribution
[†]Corresponding author

Figure 1: An overview of the MM-MATH benchmark design. The problems are classified along their difficulty, grade level, and knowledge point. We include both outcome evaluation and process evaluation to identify and attribute the error in model's reasoning process.

reasoning capabilities of LMMs. For difficulty, we classify problems into three levels—easy, medium, and hard—based on the accuracy of human students on the problems. For grade level, we include problems in middle school, encompassing all relevant visual math problems taught in each grade. For knowledge points, each problem is classified according to a predefined three-level knowledge taxonomy by experienced teachers. These comprehensive annotations in the MM-MATH dataset result in clear difficulty distinction, extensive data coverage, and systematic knowledge organization.

We conduct an extensive evaluation of both open-source and closed-source LMMs on MM-MATH. Outcome evaluation reveals that our benchmark poses significant challenges for existing LMMs. For example, the latest Sota model, GPT-4o (OpenAI, 2024), achieves an accuracy of only 31%, compared to an 82% accuracy of human students. Moreover, all models perform poorly on hard-level problems, with none exceeding 11% accuracy, and some models even fail to solve any problems correctly. We further find that current LMMs' multimodal reasoning remains primarily text-based, lacking effective utilization of graphical information. This is evidenced by the minimal accuracy difference—only 2-3 percentage points—between when the model is given only textual input and when it is provided with both text and images. Our process evaluations show that diagram misinterpretation accounts for more than 50% of the total errors for current LMMs, suggesting the most critical direction for improvement is enhancing their

abilities to recognize and interpret math diagrams.

2 MM-MATH

2.1 Overview of MM-MATH

Design Principle. Multimodal mathematical reasoning tasks demand an understanding of both the problem's text and the associated diagram, requiring math reasoning to produce a step-by-step solution that leads to the final answer. We adopt an open-ended format for two reasons: 1) Other formats, such as multiple choice, make it easier for the model to guess the correct answer by chance (Wang et al., 2024b). 2) Open-ended format better facilitates step-by-step solution process to help identify the error in the model's response. We adhere to the following principles when constructing MM-MATH:

- **Comprehensive coverage**: We aim to cover as many types and difficulty levels of problems as possible. Consequently, we collect all math problems that contain visual content from exams and textbooks used in secondary schools.

- **Computation problems only**: While math problems may include proofs, computations, and drawings, we exclusively select computation-type problems for our dataset.

- **Uniform data format**: Each problem in the dataset includes a question statement, an image, a human-annotated step-by-step solution process, and multi-dimensional metadata annotations.

- **Multi-dimensional metadata annotations**: For each problem, we also provide its grade level, dif-

Benchmark	Size	Question Type	Grade	Fine-grained Classification	Process Evaluation
UniGeo (Chen et al., 2022)	4,998	choice	middle school	✓	
GeoQA (Chen et al., 2021)	5,010	choice	middle school		
GeoQA+ (Cao and Xiao, 2022)	2,518	choice	middle school		
Geometry3K (Lu et al., 2021)	3,002	choice	middle school		
OlympiadBench (He et al., 2024)	3,102	open-ended	Olympiad-level	✓	
MathVista (Lu et al., 2023)	6,141	choice & open-ended	-		
MathVerse (Zhang et al., 2024)	2,612	choice & open-ended	-		✓
MM-MATH	5,929	open-ended	middle school	✓	✓

Table 1: Comparison of our MM-MATH benchmark with existing multimodal benchmarks. For the 'size' column, we only include the number of multimodal math problems in each benchmark.

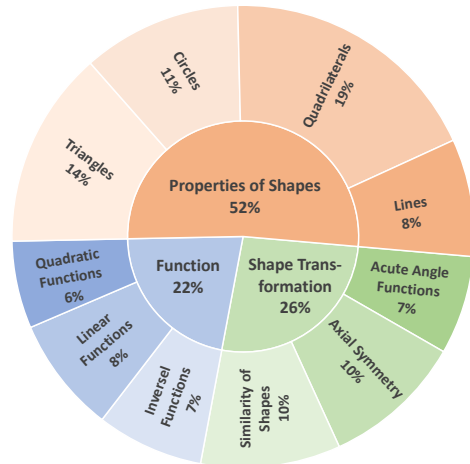| Statistic	Number
Total Problems	5,929
Unique Number of Images	4,623
Difficulty	
*Easy	378
*Medium	4,488
*Hard	1,063
Grade	
*Grade Seven	682
*Grade Eight	2,590
*Grade Nine	2,657
Average Question Length	488
Average Answer Length	275
Max Question Length	2,391
Max Answer Length	2,781

Table 2: Key statistics of MM-MATH.

Figure 2: Knowledge point distribution of MM-MATH. *Properties of Shapes* refers to the characteristics of different shapes, *Shape transformation* investigates the deformation and movements of shapes, and *Function* refers to the mutual reasoning between algebraic expressions and graphs.

ficulty, and knowledge point tagging from human educational taxonomy as its metadata.

Dataset Overview. MM-MATH is the first multimodal math benchmark to include process evaluation and fine-grained classification, as highlighted in the comparison of existing multimodal benchmarks in Table 1. Detailed statistics for MM-MATH are provided in Table 2, and the distribution of knowledge points is illustrated in the pie chart in Figure 2.

2.2 Dataset Construction Pipeline

Data collection. The problems in MM-MATH dataset are sourced from the 21st Century Education Network[1], which is one of the largest online question banks for primary and secondary schools in China. It provides a comprehensive collection of challenging, curriculum-aligned, and exam-relevant questions designed to assess student learning capabilities. We restrict the problems from the 2021-2022 academic year, manually filtering for computational math problems with visual context. Further details on the construction of MM-MATH, including grade level and year, are provided in Appendix J.

Format transformation. Most of the problems from the original database are in MathML format. However, considering the widespread use of LaTeX in existing mathematical datasets, we devise a systematic approach to convert MathML into standard LaTeX format for easier integration with other datasets. Specifically, we utilize MathConverter[2] to transform MathML representations of mathematical formulas into LaTeX. For instance, $\frac{1}{2}$ in MathML is converted to \frac{1}{2} in LaTeX. Additionally, we establish string conversion rules to change symbol elements into LaTeX format. For example, we convert "\text{△}" to "\tri-

---

[1] We have obtained official authorization from 21st Century Education for the publication of the MM-MATH dataset.

[2] https://github.com/hexinnovation/MathConverter.

angle". To address the use of non-standard punctuation in Chinese strings, such as full-width plus signs, we leverage GPT-4 (Achiam et al., 2023) for conversion, with manual verification of the final output. This systematic process ensures the accuracy of numerical values in LaTeX while maintaining readability and standardization in the output. During GPT-4 processing, we also encapsulate the final answers within \boxed{}, a technique inspired by the construction approach of the MATH dataset (Hendrycks et al., 2021), which facilitates comparison with groundtruth answers for outcome evaluation.

Our collected data contains four distinct question types: multiple-choice, fill-in-the-blank, open-ended, and composite questions. We convert them into uniform open-ended questions in the following manners. For multiple-choice and fill-in-the-blank questions, we rephrase them into open-ended forms and extract their explanations as step-by-step derivations. For composite questions with a common textual problem and multiple sub-questions, we treat each statement as the premise for sub-questions, integrating the conclusions of preceding sub-questions into the subsequent ones. More details for the transformation process are presented in Appendix B.

Additionally, since the original data is in Chinese, catering to Chinese students, we translate the dataset into English using GPT-4. This effort aims for a fairer comparison of LMMs trained in different languages. During our evaluation, we observed that some LMMs exhibit poor comprehension of Chinese content. Moreover, when handling Chinese content containing LaTeX, produce garbled and unstable outputs, whereas switching to English results in stable outputs. For this reason, we opted to use English to better accommodate the characteristics of mainstream LMMs. To further ensure translation accuracy, we validated the process by sampling 1,000 problems, finding that 988 translations were correct. The remaining errors were attributed to inaccuracies in the GPT-4 translation, which were subsequently corrected.

Fine-grained classification. We categorize our dataset across several dimensions, including difficulty, grade level, and knowledge point. Problems are classified by difficulty—simple, medium, and hard—based on the average accuracy achieved by students. Simple problems have a scoring rate above 85%, medium between 70% and 85%, and hard below 70%. From Table 2, it can be seen

that the number of problems of each difficulty level follows a Gaussian distribution.

Next, we organize questions by educational grade: seven, eight, and nine grade, representing the three years of junior middle school in China. Since higher-grade knowledge generally requires an understanding of lower-grade knowledge as a prerequisite, this classification allows us to better study whether the LMMs exhibit a similar dependency on prior knowledge when solving problems.

Additionally, each problem is tagged with specific knowledge points, identified based on insights from teachers. This enables targeted retrieval, application, and analysis of the model's knowledge gaps in specific areas. In Figure 2, we present the knowledge point taxonomy and the proportion of data in each category.

3 Evaluation

3.1 Evaluation Protocols

Recent advancements in LMMs have enabled the generation of textual responses for mathematical problem-solving (Chen et al., 2024; Liu et al., 2024b; Hong et al., 2023; Qi et al., 2024), a process that imitates human reasoning in mathematics. This capability introduces new evaluation criteria focusing on the generative nature of LMMs, especially concerning the intermediate solving steps. Accordingly, we propose a systematic method for assessing the performance of LMMs in the MM-MATH datasets in Figure 1, divided into three phases: (1) LMM generates formatted solutions to math problems, (2) Compare the generated solution against the groundtruth solution, and (3) Score the result to evaluate model performance and identify process errors. Specifically, we input the textual problem and associated images, prompting the model to generate solutions with answers encapsulated in \boxed{}. Our prompt details are provided in Appendix C. During the outcome comparison, we extract the final results from \boxed{}. For process comparison, we use GPT-4V (OpenAI, 2023) to automatically perform a comparative analysis of the model-generated solutions against the groundtruth solutions.

3.2 Evaluation Strategy

Existing large model math benchmarks (He et al., 2024; Wang et al., 2024a; Liu et al., 2024c) predominantly use binary comparison to assess the problem-solving capabilities of LLMs or LMMs,
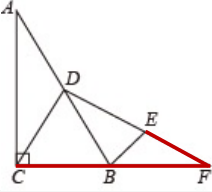
Response:

First, since CD is the median, point D is the midpoint of AB. so AD=AB.

Moreover, we know $BF = \frac{1}{2}BC$, we have CF = 2BC = 8.

…

Next, we observe that BCFE is a parallelogram, …, thus $AC = \sqrt{12} = \sqrt{3}$. Therefore, the final answer is $\sqrt{3}$.

Question: … and CB is extended to point F such that BF = BC. … what is the length of AC?

Answer: 5

Process evaluation

II. Reasoning Error
- Should be AD=DB

IV. Textual condition misunderstanding
- Should be BF=BC

I. Diagram misinterpretation
- Not a parallelogram

III. Calculation error
- $AC = \sqrt{12} = 2\sqrt{3}$

Figure 3: Example for four different types of errors in multimodal math reasoning.

focusing on the final answer as the primary indicator of the model's capability. However, the accuracy of an answer typically relies on a correct intermediate reasoning process, with accurate reasoning steps leading to correct answers, and incorrect reasoning leading to erroneous solutions. By analyzing the model's solution process, we can identify the causes of errors and provide a more accurate assessment. Thus, our evaluation incorporates two methods: outcome evaluation and process evaluation, designed to assess both the final answer and the model's reasoning process.

Outcome Evaluation.

Our evaluation requires that answers generated by LMMs be encapsulated within \boxed{}, enabling direct comparison. We judge final answers according to their category: (1) For numerical answers, we accept the model's answers as long as the numerical gap to the groundtruth answer falls within a permissible error margin, e.g., 1.414 is acceptable for an answer of $\sqrt{2}$ as their difference is less than 0.01. (2) For expression-type answers such as $y = ax + b$, we utilize the SymPy package to simplify expressions. We then compare the model's simplified output with the groundtruth expression for exact matching. (3) For interval-type answers like (a, b) or $a < x < b$, we standardize them into the format (a, b) and verify the equality of boundary values. Additionally, we address special cases where models append extra signs to final results (e.g., cm) or generate exponential values like 2^{2024}, by removing the extra sign and transforming the values for proper comparison. We manually verified 500 evaluation results using our outcome evaluation pipeline and found only 13 errors.

Process Evaluation. The problem-solving process of the multimodal model involves multiple factors, including a deep understanding of the problem conditions, extracting information from diagrams, and utilizing the models' knowledge to derive results. Consequently, our process evaluation takes the original textual question, associated image, and the groundtruth solution, and uses GPT-4V to compare the content generated by LMMs, with the prompt shown in Appendix C. The solutions generated by LMMs may contain numerous errors. In our prompt design, we aim to identify the **first error** in the model's generated process compared to groundtruth, since it is often the initial error that leads to further mistakes, resulting in incorrect outcomes. We use this first error to classify the cause of error in our process evaluation. Through deeper examination, we find that the first identified error may sometimes not be the main error of the models' solution, which we will analyze further in Appendix D. We classify the errors into four types, exemplified in Figure 3.

I. *Diagram misinterpretation*: This refers to the LMM's inability to accurately understand the elements and their attributes in diagrams, such as the shapes, geometries, and their spatial relationships.

II. *Reasoning error*: This occurs when the model lacks or incorrectly applies logical reasoning knowledge. For instance, in the case of Figure 3, the model incorrectly reasons that $AD = AB$ from D is the midpoint of AB, while $AD = DB$ should be the correct deduction.

III. *Calculation error*: This error arises from the computational step during problem-solving and includes mistakes caused by miscalculations in equations and functions.

IV. *Textual condition misunderstanding*: This type of error involves a model misinterpreting the given conditions of a textual problem. For example, in Figure 3, the problem states that $BF = BC$, but the model mistakenly interprets this condition as $BF = \frac{1}{2}BC$ during the solution process.

Model	Easy	Medium	Hard	Seven	Eight	Nine	Trans	Shape	Func	Average
Baseline										
Human	90.7	81.9	47.6	85.6	73.7	77.9	81.1	83.2	77.5	80.4
Large Multimodal Models (w/o Image)										
Gemini-Pro-V	10.1	5.7	1.8	10.0	5.3	6.7	6.6	5.7	6.4	6.2
Claude-3-Opus	31.7	17.3	7.2	32.5	14.9	2.2	20.8	18.5	12.9	19.2
GPT-4	37.0	20.3	7.2	38.7	17.1	26.2	23.3	21.4	18.1	22.5
GPT-4V	35.2	18.1	7.2	31.2	17.2	22.3	18.4	21.4	13.3	20.4
GPT-4o	41.4	23.9	3.6	35.0	23.9	30.5	22.8	29.7	19.4	27.6
Large Multimodal Models (w/ Image)										
DeepSeek-VL-7B-Chat	17.4	4.7	1.4	7.5	6.6	3.9	3.4	6.0	3.5	5.4
Yi-34B-Chat	12.9	5.0	1.5	21.3	5.6	3.5	5.0	7.6	3.8	6.5
LLaVA-V1.6-34B	8.8	5.4	1.8	12.6	6.5	4.2	4.0	6.5	3.8	5.8
InternVL-4B-Chat-1.5	18.5	10.7	1.8	12.5	11.1	11.9	11.4	12.3	5.5	11.6
Qwen-VL-Max	14.5	11.2	3.6	16.2	1.1	11.3	11.0	12.5	10.5	11.4
Gemini-Pro-V	19.3	8.2	0.0	1.5	7.4	11.5	10.4	10.6	7.1	9.7
Claude-3-Opus	29.5	19.3	3.6	32.5	16.4	23.0	20.6	21.7	16.9	20.3
GPT-4V	37.8	21.2	1.8	28.7	17.9	28.0	22.2	24.7	19.5	23.1
GPT-4o	**45.8**	**30.0**	**10.9**	**40.0**	**26.0**	**36.0**	**30.7**	**33.7**	**26.2**	**31.8**

Table 3: The outcome performance of both closed-source and open-source large models on MM-MATH in comparison with the human-level baseline. The evaluation involves three dimensions: *difficulty*, *grade levels*, and *knowledge points*, each comprised of three fine-grained classes. The results are presented as percentages of accuracy.

4 Experiments

4.1 Experimental Setup

To comprehensively investigate the challenges of MM-MATH and the mathematical proficiency of models, we structure our experiments around two setups: (1) Text-Only Reasoning and (2) Multimodal Reasoning. For the first setting, we evaluate LMMs, including Gemini-Pro-V (Gemini, 2023), Claude-3-Opus (Anthropic, 2024), GPT-4[3] (Achiam et al., 2023), GPT-4V (OpenAI, 2023), and GPT-4o (OpenAI, 2024) by providing only the textual contexts (*i.e.,* questions) as inputs. For the second setting, we feed the entire multimodal contexts (*i.e.,* questions and images) as inputs and evaluate both closed-source LMMs including GPT-4V, GPT-4o, Claude-3-Opus, Qwen-VL-Max (Bai et al., 2023) and open-source LMMs for DeepSeek-VL-7B-Chat (Lu et al., 2024), Yi-34B-Chat (Young et al., 2024), InternVL-4B-Chat-V1.5 (Chen et al., 2024), and LLaVA-V1.6-34B (Liu et al., 2024a).

All selected models are capable of generating responses in the expected format, thus ensuring the validity of the evaluation.

4.2 Outcome Evaluation Results

We first analyze the performance of all models on the final outcomes of MM-MATH in comparison to a human-level baseline (the average performance

of middle-school examinees from the online platform). The experimental results are shown in Table 3. Here are our main findings from the results.

MM-MATH presents substantial challenges for current LMMs From the evaluation results, we find that the most representative closed-source model to date, GPT-4o, performed the best across the board, achieving an average accuracy of 31.8%, which significantly outperformed the best open-source model, InternVL-4B-Chat-1.5, with an average accuracy of 11.6%. However, compared to the human-level baseline of 80.4%, this best performance of the LMM still remains substantial room for improvement by 48.6%.

LMMs gain limited benefits from visual contexts Another notable observation is that LMMs with the text-only setups (*i.e.,* only questions as inputs) exhibit only slight degradation in performance compared to the multimodal setups (*i.e.,* questions and images as inputs). For example, there are differences of 4.2%, 2.7%, and 0.8% for the models GPT-4o, GPT-4v, and Claude-3-Opus, respectively. This result suggests that current LMMs primarily rely on linguistic knowledge to solve mathematical problems, and their utilization of visual contexts is limited. Detailed case studies are provided in Appendix E. Further analysis of GPT-4o's performance reveals that it can infer some certain answers without visual input. Selected cases are detailed in

---

[3]We use the `gpt-4-0125-preview` version for GPT-4.

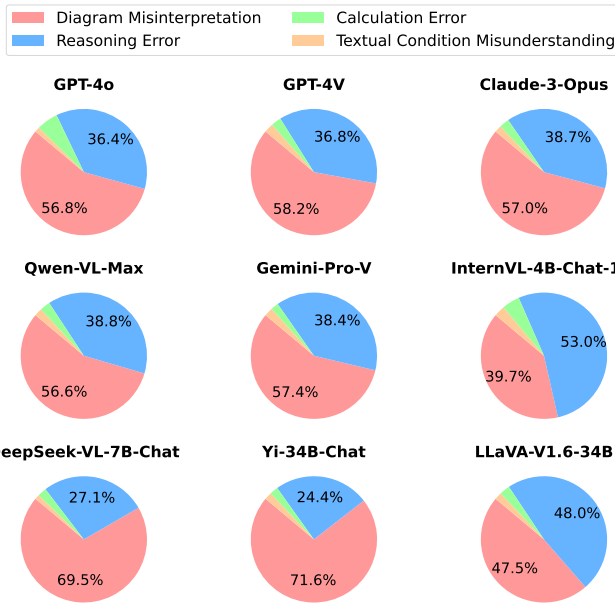Figure 5: Number of the first two errors in evaluated LMMs.

Figure 4: Proportion of four types of errors in various LMMs, with diagram misinterpretation errors and reasoning errors constituting the majority.
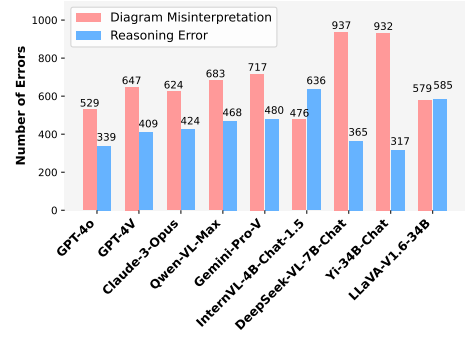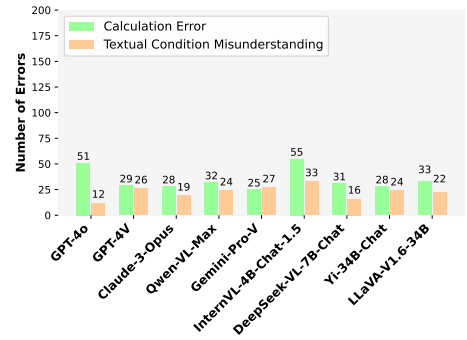
Figure 6: Number of the last two errors in evaluated LMMs.

Appendix F.

Conclusion from discriminative evaluation dimensions and capability distribution In the difficulty dimension of MM-MATH, we can see the discriminative stepwise degradations in models performance on progressively challenging subsets (*e.g.,* the 10.9%, 30.0%, and 45.8% accuracy scores on *Easy*, *Medium* and *Hard* subsets for GPT-4o). This result indicates that the proposed evaluation dimensions exhibit a significant differentiation across three difficulty levels, making it more beneficial for exploring the capability shortcomings of models. In addition, the three types of knowledge points also provide us with opportunities to understand the capabilities of models from different fine-grained perspectives.

Regarding the evaluation results on different grade levels, one notable finding is that the accuracy distribution of most models across the three grade levels is similar to the distribution of human behaviors. For example, the models GPT-4o, GPT-4V, and Claude-3-Opus all showed the best performance on the Seventh-grade subset (with 40.0%, 28.7%, and 32.5% accuracy scores), followed by

the ninth-grade subset (with 36.0%, 28.0% and 23.0% accuracy scores), and were the least accurate on the seventh-grade subset (with 26.0%, 17.9% and 16.4% accuracy scores, respectively). This result suggests that the learning curve of LMMs in solving mathematical problems is similar to that of humans but falls short of reaching the human cognitive level.

4.3 Process Evaluation Results

Benefiting from the comprehensive annotation, we further evaluate the models performance on the solution process to thoroughly investigate the causes of errors and pinpoint the weaknesses of LMMs. Considering the variability in natural language expressions, we employ GPT-4V to compare the solutions generated by LMMs with the groundtruth solutions, and identify the first error in the solutions to analyze the causes of errors. We empirically find that this method can effectively align the solutions from LMMs with the groundtruth, enabling an unbiased validation of the errors. Though effective, we find that there is still room for improvement in this measurement, with approximately 9% of errors not being correctly identified (see the detailed

1364

analysis in Appendix G).

Figure 4 illustrates the proportion of different error types in both open-source and closed-source LMMs. Figure 5 and 6 further show the number of errors for each error type. Our main findings are detailed below.

Weak comprehension of elements in images is a major cause It is evident that errors related to the recognition of image elements or their attributes constitute the highest proportion, exceeding half of the total errors. This indicates that existing LMMs cannot yet sufficiently incorporate image information into their reasoning processes, limiting their efficacy in multimodal reasoning. Intriguingly, among closed-source LMMs—GPT-4o, GPT-4V, Claude-3-Opus, Gemini-Pro-V, and Qwen-VL-Max—the proportion of errors in image recognition are highly consistent, around 57%. This might imply that the visual encoder modules used by these models have common issues and cannot handle certain types of images. Additionally, the much lower proportion of diagram misinterpretation errors in InternVL-4B-Chat-1.5 (39.7%) explains why a 4B small model has even better overall performance than Gemini-Pro-V (57.4%) or Qwen-VL-Max (56.6%). Therefore, the key to enhancing the LMM's multimodal math problem-solving ability lies in understanding the visual context, and this step does not necessitate a large model size. Examples of reasoning errors involving image elements and attributes are provided in Appendix I.

Multimodal models exhibit poor use of theorems during reasoning We find that reasoning errors in large language models (LLMs) are often due to the incorrect application of theorems, accounting for about 40% of overall errors. Misuse or omission of theorems misleads these LMMs, leading to errors (e.g., GPT-4V misuses the cosine rule, resulting in no solution, as detailed in Appendix H). Unlike image understanding, we find that a larger model size effectively helps reduce reasoning errors in the model. For instance, while InternVL-4B-Chat-1.5 exhibits fewer image understanding errors even with smaller model size, it still encounters more reasoning errors (636) compared to larger models such as Gemini-Pro-V (480) and Qwen-VL-Max (468).

Calculation is not a primary issue but reflects a capability gap In the process evaluation of LMMs, calculation errors constitute a relatively lower proportion. However, the error in some models (e.g., GPT-4o, 51 errors) is significantly higher compared to others (e.g., GPT-4V, 29 errors). This indicates that while calculation is not the primary problem, equipping them with more powerful numerical computation capabilities can further boost the models' problem-solving success rates.

Models have an effective understanding of the textual problem As shown in Figure 4, among all nine models from both open-source and closed-source, the proportion of errors due to misunderstanding of the textual conditions is extremely small (less than 2% of the total errors). This suggests that the text-based capabilities of LMMs are not the bottleneck in solving multimodal mathematical problems. Instead, we should focus more on fine-grained recognition and reasoning of visual content to enhance the capabilities of LMMs.

5 Related Work

Using large models to solve mathematical problems has recently become a research hotspot. GSM8k (Cobbe et al., 2021) has widely been used to evaluate the mathematical abilities of various LLMs (Touvron et al., 2023; Anil et al., 2023; Gao et al., 2023b). However, its problems are relatively simple, and many models can achieve an accuracy rate of 90% or higher. Recently, more challenging mathematical benchmarks (Hendrycks et al., 2021; Liu et al., 2024c; He et al., 2024) have emerged to further advance mathematical reasoning in language models, but these are typically text-only based reasoning.

Multimodal mathematical benchmarks trace back to the study of geometry problems (Seo et al., 2015; Chen et al., 2022), where geometric elements are described through a specialized parsing language (Seo et al., 2015; Zhang et al., 2022; Hao et al., 2022) or text described language (Gao et al., 2023a). Recent rapid developments in LMMs (Alayrac et al., 2022; Wang et al., 2023; Liu et al., 2024b; Qi et al., 2024) have led to numerous multimodal math benchmarks (Lu et al., 2023; Yue et al., 2024; Ying et al., 2024) to assess their capabilities. However, these benchmarks primarily composed of multiple-choice questions, evaluating model performance based on outcome examination. Given the dual nature of multimodal models—integrating both images and text—such simplistic evaluations are inadequate. Although some benchmarks, like MathVerse (Zhang et al., 2024), have

begun to focus on the problem-solving process, they still rely on a binary evaluation approach. In comparison, our MM-MATH benchmark is constructed with step-by-step solution which enables both outcome and process evaluations of LMMs.

6 Conclusion

This paper introduces MM-MATH, a challenging benchmark for evaluating multimodal math reasoning in LMMs. Our findings reveal while current LMMs demonstrate some reasoning ability, they heavily rely on textual information and struggle to utilize visual cues. This is evidenced by the minimal accuracy difference between text-only and multimodal settings, and the prevalence of diagram misinterpretation errors. MM-MATH's fine-grained classification highlights the need for models that can handle varying problem difficulties and leverage knowledge across different grade levels.

7 Limitations

We limit our benchmark's mathematical knowledge to the middle school level, representing only a portion of K-12 education. In the future, we plan to expand the scope of MM-MATH to include high school and college-level multimodal mathematics. Our evaluation results highlight the current deficiencies of LMMs in solving mathematical problems. While improvements to LMMs have not yet been made to address these shortcomings, our next step involves targeted training to enhance the models' problem-solving capabilities. We believe our dataset will significantly aid this process, as they contain detailed solutions paired with each problem.

8 Acknowledgements

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Anthropic. 2024. Claude3 system card.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Jie Cao and Jing Xiao. 2022. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1511–1520.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. 2023a. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.

Team Gemini. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Yihan Hao, Mingliang Zhang, Fei Yin, and Lin-Lin Huang. 2022. Pgdp5k: A diagram parsing dataset for plane geometry problems. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1763–1769. IEEE.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. 2023. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024c. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. *arXiv preprint arXiv:2405.12209*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786.

OpenAI. 2023. GPT-4V(ision) system card.

OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. 2024. Cogcom: Train large vision-language models diving into details through chain of manipulations. *arXiv preprint arXiv:2402.04236*.

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1466–1476.

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

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*.

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

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi:

Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Ming-Liang Zhang, Fei Yin, Yi-Han Hao, and Cheng-Lin Liu. 2022. Plane geometry diagram parsing. *arXiv preprint arXiv:2205.09363*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A Data Source for Human Performance

The 21st Century Education Network provides academic proficiency reports that analyze students' knowledge mastery after each exam. We compile the end-of-term exam scores for each problem.

B Open-Ended Transformation

Our initial collection of MM-MATH problems includes four types: multiple-choice, fill-in-the-blank, open-ended, and composite questions. For multiple-choice and fill-in-the-blank questions, which include an answer and a step-by-step solution, we modifies the final part of the questions into descriptive language, removing single choice or fill-in-the-blank answers, and using the step-by-step solution as the answer described in Figure 7. For composite questions, we treat the main textual problem as the common stem for sub-questions and used the conclusion of one sub-question as the textual problem for the next described in Figure 8.

Choice Question or Fill-in-the-Blank:	As shown in the figure, given that the diameter of circle $\odot O$ is 4 and $\angle ACB = 45°$, the length of AB is ()?
Option:	A. 2, B. $2\sqrt{2}$, C. $2\sqrt{3}$, D. $\frac{1}{2}$
Explanation:	Solution: Connect OA and OB, as shown in the figure, $\because \angle AOB = 2\angle ACB = 2\times45° = 90°$, $\therefore \triangle AOB$ is an isosceles right triangle, $\therefore AB = \sqrt{2}OA = 2\sqrt{2}$. Therefore, $AB = 2\sqrt{2}$. Hence the answer is B or $2\sqrt{2}$
Open-End Answer:	As shown in the figure, given that the diameter of circle $\odot O$ is 4 and $\angle ACB = 45°$. What is the length of AB? Solution: Connect OA and OB, as shown in the figure, $\because \angle AOB = 2\angle ACB = 2\times45° = 90°$, $\therefore \triangle AOB$ is an isosceles right triangle, $\therefore AB = \sqrt{2}OA = 2\sqrt{2}$. Therefore, $AB = \boxed{2\sqrt{2}}$.

Figure 7: An example of converting multiple-choice and fill-in-the-blank questions to open-ended format. The final part of the textual problem "()" is rewritten in descriptive language, and the main content of the explanation is used as the answer.

C Prompt Design

Table 4 details the construction of the two types of prompts. For process evaluation prompts, our repeated experiments highlighted several key points: 1) Use the term "*incorrect*" for *textual condition misunderstanding* to help GPT-4V classify the errors accurately. 2) Use the term "*misinterpretation*" for *diagram misinterpretation* errors to identify recognition mistakes during comparisons. 3) For reasoning errors, it is important to include specific examples.

For prompts that instruct the model to generate answers, we ensure the model produces a final answer enclosed in \boxed{}.

D First Error Identified

The first error identified by GPT-4V, when comparing the problem-solving process generated by LMMs to the ground truth, may not necessarily be the initial error in the problem-solving process. As shown in Figure 9, the first error determined by GPT-4V is $\triangle ABD \sim \triangle CBE$ rather than intial error

$$\frac{AC}{BC} = \frac{AD}{DE}.$$

Phase	Input	Prompt
Process Evaluation (GPT-4V)	Model's response Question Diagram Groundtruth Answer	Based on the given question stem, the diagram, and the correct answer, compare the model's response to identify the first error in model's response. Then determine which of the following categories the error belongs to, or if there is no error, classify it as category five: 1. Misinterpretation of diagram elements or properties: For example, incorrect coordinate recognition, identifying parallel lines as intersecting lines, or inventing or misusing elements or properties not present in the diagram (e.g., identifying a shape as a square when it is not). 2. Incorrect application of math theorems: For instance, wrongly applying a specific theorem, such as using the Pythagorean theorem on a non-right triangle, or omitting necessary theorems, such as failing to apply the similarity theorem to obviously similar triangles. 3. Calculation errors: Such as mistakes in addition, subtraction, multiplication, division, or square root calculations. 4. Incorrect use of given question stems: For example, if the stem states AB=1/2CD but the model generates AB=CD, indicating a failure to use the condition correctly. 5. Other: No errors. Provide a detailed analysis, including the first mistake, the reason for the classification, and the correct approach to solving the problem. If there are no errors, only provide the analysis. The output format should be: –**First error**: –**Error category**: –**Detailed analysis**:
Answer Generation (LMMs)	Question Diagram	Solve the following mathematics problem, write out the solution process according to the question, and use the same LaTeX format as the question in the solution process. Please display the final answer in the format \boxed{}.

Table 4: This table presents the prompts used for process evaluation and answer generation by various LMMs in the MM-MATH benchmark.

Composite Textual Question	As shown in the figure, in △ ABC , ∠ACB = 90°, and CD is the altitude to side AB. Fold side AC in half, and the fold line is EF . Connect CE . CD bisects ∠BCE.
Question 1:	Connect DF, prove that AF = DF.
Question 2:	Find the measure of ∠A.
Explanation 2:	Solution: ∵ EF is the axis of symmetry of AC, ∴FA = FC , EA = EC, ∴ ∠ECA = ∠A . ∵CD is the altitude to side AB, ∴ ∠CDE = ∠CDB = 90°. ∵CD bisects ∠BCE, ∴ ∠DCE = ∠DCB . Also, ∵ CD = CD, ∴△CDE ≅△CDB (ASA). ∴ ∠CED = ∠CBD. ∵ ∠CED = ∠A + ∠ECA, and ∠ECA = ∠A. ∴ ∠CBD = 2∠A. ∵ ∠ACB = 90°, ∴ ∠A + ∠B = 90°, ∴3∠A = 90°, ∴ ∠A = 30°.
Composite Question	As shown in the figure, in △ ABC , ∠ACB = 90°, and CD is the altitude to side AB. Fold side AC in half, and the fold line is EF . Connect CE . CD bisects ∠BCE. Connect DF, AF = DF. Find the measure of ∠A.
Open-End Answer:	∵ EF is the axis of symmetry of AC, ∴FA = FC , EA = EC, ∴ ∠ECA = ∠A . ∵CD is the altitude to side AB, ∴ ∠CDE = ∠CDB = 90°. ∵CD bisects ∠BCE, ∴ ∠DCE = ∠DCB . Also, ∵ CD = CD, ∴△CDE ≅△CDB (ASA). ∴ ∠CED = ∠CBD. ∵ ∠CED = ∠A + ∠ECA, and ∠ECA = ∠A. ∴ ∠CBD = 2∠A. ∵ ∠ACB = 90°, ∴ ∠A + ∠B = 90°, ∴3∠A = 90°, ∴ ∠A = 30°

Figure 8: An example of converting composite question to open-ended format. Since Question 1 is a proof, we exclude it. We treat the main stem as the stem of Question 2, and incorporate the conclusion of Question 1 (highlighted in red) as a new condition into the stem of Question 2.

E Text Reason First

Figure 10 and Figure 11 illustrate examples of multimodal reasoning. Regardless of whether all problem conditions are provided, multimodal models tend to rely solely on textual analytical methods, neglecting the information in the images. This approach increases the complexity of problem-solving and leads to a higher likelihood of errors.

F Case of GPT-4o Answer

Figure 12 illustrates an example where GPT-4o achieves the correct conclusion without viewing the image. Based on the problem description, although the side length is provided, the folding method is only shown in the image, making it impossible to solve the problem without visual reference.

G Case of Prompt Effectiveness

Figure 13 illustrates the effectiveness of our designed prompt. Although there is no textual description of DOB in the groundtruth problem-solving process, GPT-4V successfully identified and categorized the error. The error was classified as a mathematical theory error by GPT-4V because the solution did not apply the similarity theorem for reasoning.

H Misapplication of Theorems

Figure 14 illustrates a case where GPT-4V incorrectly applied the cosine theorem during problem-solving. In this case, the sides used in the cosine theorem do not form a triangle in the image. Despite this, the model forcibly applied the theorem, leading to calculation errors. This improper use of the theorem on incorrect image elements further resulted in erroneous problem-solving steps.
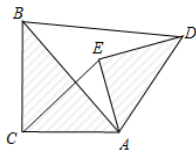
Question: As shown in the figure, in $\triangle ABC$ and $\triangle ADE$, $\angle ACB=\angle AED=90°$, $\angle ABC=\angle ADE$. Connect BD and CE. If AC:BC=3:4, what is the ratio BD:CE?

Solution: Since $\angle ACB=90°$ and AC:BC=3:4, we have $\frac{AB}{AC}=\frac{5}{3}$. $\because \angle ACB=\angle AED$ $=90°$ and $\angle ABC=\angle ADE$, it follows that $\triangle ABC \sim \triangle ADE$, thus $\frac{AB}{AD}=\frac{AC}{AE}$ and $\angle DA$ $E=\angle BAC$. Therefore, $\angle DAB=\angle EAC$, and $\triangle DAB\sim\triangle EAC$, hence $\frac{BD}{CE}=\frac{AB}{AC}=\boxed{\frac{5}{3}}$.

First Error

GPT-4V Answer: From the problem statement, we know $\triangle ABC \sim \triangle ADE$ because they are both right triangles and share a common acute angle. According to the properties of similar triangles, we have $[\frac{AC}{BC}=\frac{AD}{DE}]$ .. Therefore, $\triangle ABD \sim \triangle CBE$. Thus, we have $[\frac{BD}{CE}=\frac{AB}{BC}]$ ... Therefore, $[\frac{BD}{CE}=\frac{5}{4}]$ So the final answer is $\boxed{\frac{5}{4}}$.

First error: "Therefore, $\triangle ABD \sim \triangle CBE$".
Error category: 1. Misinterpretation of diagram elements or properties.
Detailed analysis: The model made a mistake in determining the similarity of $\triangle ABD$ and $\triangle CBE$. In reality, $\triangle ABD$ and $\triangle CBE$ are not similar. …

Figure 9: An example of first error detecting by GPT-4V
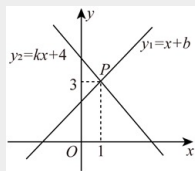
Question: As shown in the figure, the graph of the linear function $y_1 = x + b$ intersects with the graph of the linear function $y_2 = kx + 4$ at point P(1, 3). What is the solution set for the inequality $x + b \leq kx + 4$ with respect to x?

Solution: According to the graph, when $x \leq 1$, $x + b \leq kx + 4$, that is, the solution set for the inequality $x + b \leq kx + 4$ with respect to x is $\boxed{x \leq 1}$.

GPT-4V Answer: First, based on the lines $y_1 = x + b$ and $y_2 = kx + 4$ intersecting at point P(1, 3), we can derive two equations to determine the values of b and k. For $y_1 = x + b$, substituting the coordinates of point P(1, 3), we get: 3 = 1 + b Thus, we can solve for b: b = 2 For $y_2 = kx + 4$, similarly substituting the coordinates of point P(1, 3), we get: 3 = k + 4 Solving for k: k = -1 Now we know the equations of the lines are $y_1 = x + 2$ and $y_2 = -x + 4$. Given the inequality $x + b \leq kx + 4$ from the problem, substituting b = 2 and k = -1 gives: $x + 2 \leq -x + 4$ Next, solve this inequality: $2x \leq 2$, $x \leq 1$, So the solution set is x less than or equal to 1. The final answer is $\boxed{x \leq 1}$.

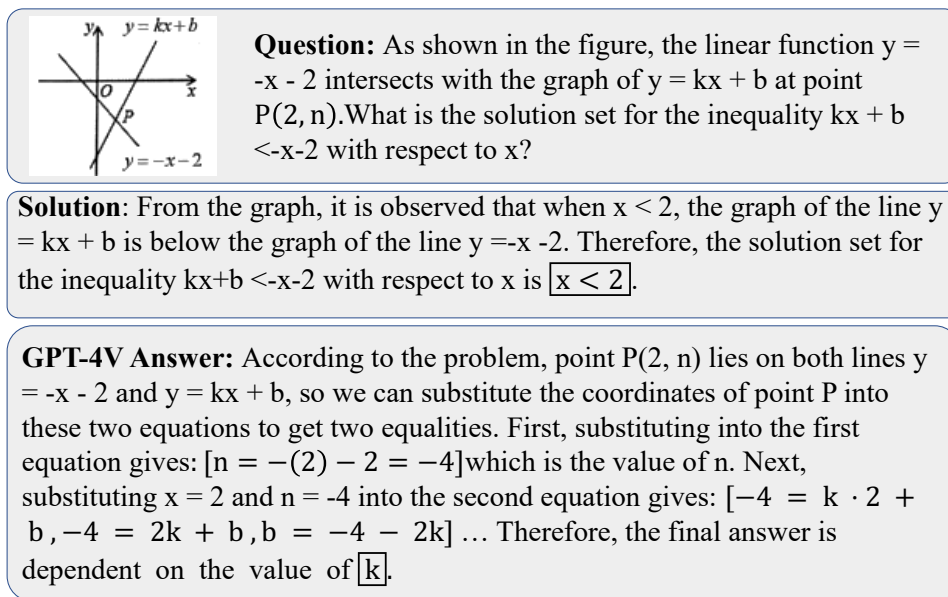Figure 10: A case where LMMs reasoning relies solely on textual analysis, resulting in increased complexity.

Question: As shown in the figure, the linear function y = -x - 2 intersects with the graph of y = kx + b at point $P(2, n)$.What is the solution set for the inequality kx + b <-x-2 with respect to x?

Solution: From the graph, it is observed that when x < 2, the graph of the line y = kx + b is below the graph of the line y =-x -2. Therefore, the solution set for the inequality kx+b <-x-2 with respect to x is $\boxed{x < 2}$.

GPT-4V Answer: According to the problem, point P(2, n) lies on both lines y = -x - 2 and y = kx + b, so we can substitute the coordinates of point P into these two equations to get two equalities. First, substituting into the first equation gives: $[n = -(2) - 2 = -4]$ which is the value of n. Next, substituting x = 2 and n = -4 into the second equation gives: $[-4 = k \cdot 2 + b, -4 = 2k + b, b = -4 - 2k]$ … Therefore, the final answer is dependent on the value of $\boxed{k}$.

Figure 11: A case where LMMs reasoning ignores image information, relying only on text and leading to errors.

Question: As shown in the figure, fold a rectangular paper ABCD as illustrated twice to form an isosceles right-angled triangle BEF. If $BC = 1$, what is the length of AB? **True Answer:** $\frac{\sqrt{2}}{2}$

GPT-4o Answer: Assume the coordinates of the rectangle paper ABCD are: $(A(0,0)), (B(a, 0)), (C(a, 1)), (D(0,1))$. First Fold: Fold the rectangle along the BC edge such that point B coincides with point C, and let the fold line be MN. … Since $BC = 1$ is a constant from the initial step, and after the transformation $AB = BC$: The final answer is $\boxed{\frac{\sqrt{2}}{2}}$
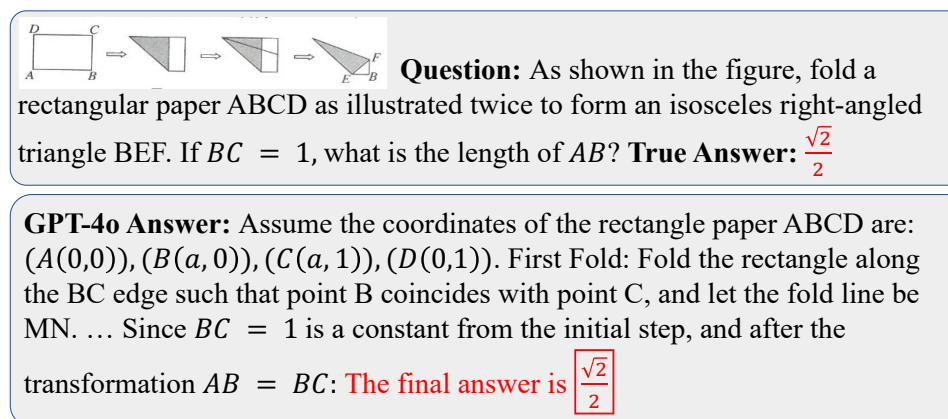
Figure 12: A case where GPT-4o generates the correct answer without the use of an image.

I Element Recognition Error in Image

We present two cases where GPT-4V exhibited errors in image element recognition during problem-solving. In Figure 15, the image contains only parallel lines; however, GPT-4V incorrectly identified a triangle. In Figure 16, AFD is a straight line, but GPT-4V mistakenly perceived it as a triangle. These cases demonstrate GPT-4V's deficiencies in accurate image element recognition, leading to erroneous reasoning.

J Dataset Construction and Problem Type Analysis

The 21st Century Education Network provides exam questions from elementary to high education. Upon analyzing elementary school math problems, we observe significant variation in problem types, many of which lack consistency. For example, tasks such as number recognition from images are more appropriate for evaluating multimodal reasoning rather than mathematical reasoning in multimodal models. For high school math problems, we will release them after fine-grained validation. The problems are collected from 2021-2022, as two years of data are sufficient to comprehensively cover all relevant problem types without introducing redundancy.
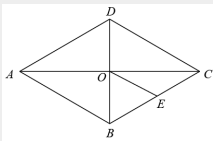
Question: As shown in the figure, the diagonals AC and BD of rhombus ABCD intersect at point O. Given that AC = 8 and BD = 6, and point E is the midpoint of BC, what is the length of OE?

Solution: ∵quadrilateral ABCD is a rhombus,∴OC=OA= $\frac{1}{2}$ AC=4,OB=OD= $\frac{1}{2}$ BD = 3,∠BOC = 90°.∴AB= $\sqrt{OA^2 + OB^2}$ =5.∵point E is midpoint of AB, and O is the midpoint of AC,∴OE is the median of △ABC,∴OE= $\frac{1}{2}$ AB= $\boxed{\frac{5}{2}}$.

GPT-4V Answer: …, We also know that since all sides of the rhombus are equal, AB = BC = CD = DA. Because OE is the median of the rhombus's diagonal BD, it bisects triangle DOB, making △BOE ∼△DOE. Now, … , we have $\frac{OE}{BO} = \frac{BE}{DO}$ .Thus, OE = $\frac{BE \times BO}{DO} = \frac{2.5 \times 3}{3}$ = 2.5. Therefore, the final answer is $\boxed{2.5}$.

First error: Because OE is the median of the rhombus's diagonal BD, it bisects triangle DOB, making △BOE ∼△DOE.
Error category: 2. Incorrect application of math theorems
Detailed analysis: The model incorrectly assumes that OE is the median of diagonal BD and incorrectly applies the properties of similar triangles. ...

Figure 13: A case demonstrating the effectiveness of our prompt. GPT-4V successfully identified the reasoning error DOB, which was not present in the solution.
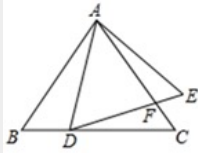
Question: As shown in the figure, △ABC and △ADE are both equilateral triangles, point D is on side BC, DE intersects AC at point F, and if AB=6, AD=5, and CD=4, what is the length of EF?

Solution: ∵△ ABC and △ ADE are both equilateral triangles, ∴ ∠B = ∠C = ∠ADE = 60°, DE = AD = 5, ∵ ∠ADC = ∠ADE + ∠CDE = ∠B + ∠BAD, ∴ ∠CDE = ∠BAD, ∴△ ABD ∼△ DCF, ∴ $\frac{AB}{DC} = \frac{AD}{DF}$, ∴ $\frac{6}{4} = \frac{5}{DF}$, ∴ DF = $\frac{10}{3}$, ∴ EF = DE − DF = $\boxed{\frac{5}{3}}$.

GPT-4V Answer: … we have BC = 6. Given that CD = 4, we find BD = BC - CD = 6 - 4 = 2. Next, applying the cosine rule to find the measure of ∠BCD: $BC^2 = BD^2 + CD^2 − 2×BD×CD× \cos ∠BCD$, Substituting BC = 6, BD = 2, and CD = 4 into the formula to solve for $\cos ∠BCD$, …, However, we obtained a result of $\cos ∠BCD$ greater than 1, which is impossible since the cosine value must be within the range [-1, 1]. This indicates that there might be an error in the problem's information.

First error: $BC^2 = BD^2 + CD^2 − 2×BD×CD× \cos ∠BCD$
Error category: 2. Incorrect application of math theorems
Detailed analysis: The model incorrectly applied the cosine rule in the solution process. The problem clearly states that $\triangle ABC$ and $\triangle ADE$ are equilateral triangles, …

Figure 14: A reasoning error case: misuse of the cosine rule
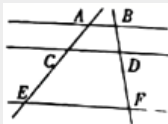
Question: As shown in the figure, given AB ∥ CD ∥EF and BD: DF = 2: 5, what is the value of $\frac{AC}{AE}$?

GPT-4V Answer: According to the figure, segments AB ∥CD ∥EF and BD:DF = 2:5. Since AB ∥ CD, by the properties of similar triangles, we know that triangles ABD and ACD are isosceles, thus AD = BD. Similarly, since CD ∥EF, triangles CDE and EDF are isosceles, so ED = DF. ,…

Figure 15: A case of diagram misinterpretation : parallel lines misidentified as a triangle.
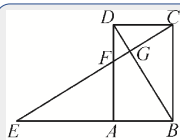
Question: As shown in the figure, quadrilateral ABCD is a rectangle, point E is on the extension of BA, AE = AD, and EC intersects AD and BD at points F and G, respectively. If AF = AB, what is the value of $AD : AB$?

GPT-4V Answer: First, since ABCD is a rectangle, $AB \perp AD$, and AB = CD and AD = BC. According to the problem, AE = AD, so △ AED is an isosceles right triangle, thus $∠AED = 45°$.Additionally, since AF = AB = CD,△ AFD and △ BCD are congruent, $so∠ADF = ∠BDC = 90°$.Therefore, …

Figure 16: A case of diagram misinterpretation: a straight line misidentified as a triangle.