Are MLLMs Robust Against Adversarial Perturbations? RoMMath: A Systematic Evaluation on Multimodal Math Reasoning

Yilun Zhao* **Guo Gan*** **Chengye Wang*** **Chen Zhao** **Arman Cohan**

Yale NLP RoMMath Team

https://github.com/yale-nlp/RoMMath

Abstract

We introduce RoMMath, the first benchmark designed to evaluate the capabilities and robustness of multimodal large language models (MLLMs) in handling multimodal math reasoning, particularly when faced with adversarial perturbations. RoMMath consists of 4,800 expert-annotated examples, including an original set and seven adversarial sets, each targeting a specific type of perturbation at the text or vision levels. We evaluate a broad spectrum of 17 MLLMs on RoMMath and uncover a critical challenge regarding model robustness against adversarial perturbations. Through detailed error analysis by human experts, we gain a deeper understanding of the current limitations of MLLMs. Additionally, we explore various approaches to enhance the performance and robustness of MLLMs, providing insights that can guide future research efforts.

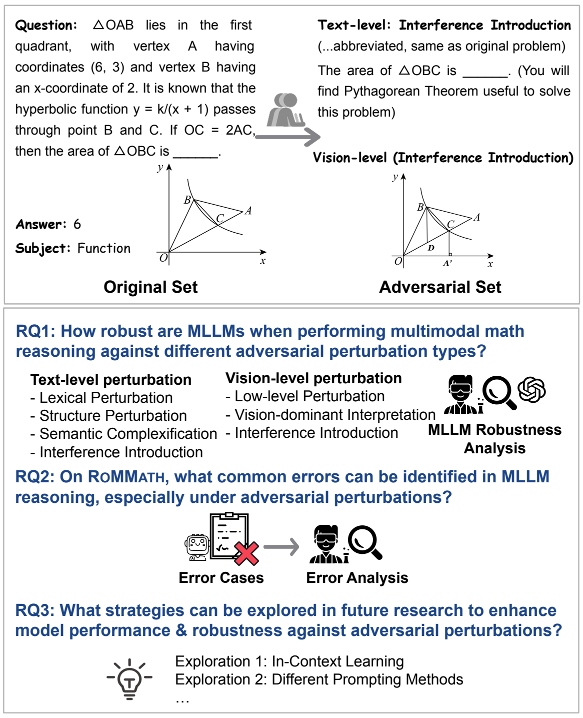

Figure 1: Overview of this research. (Top) An illustration of RoMMath benchmark construction; (Bottom) the three research questions investigated in this paper.

1 Introduction

Multimodal math reasoning is a compelling area for assessing the reasoning capabilities of MLLMs because it involves complex tasks that require accurate interpretation and reasoning across both visual and textual modalities (Chen et al., 2021; Masry et al., 2022a; Lu et al., 2024b; Zhang et al., 2024b; Wang et al., 2024a; Chen et al., 2024a; Liang et al., 2024a). Recently-released MLLMs have shown remarkable performance on various multimodal math reasoning benchmarks (Lu et al., 2024a; Liu et al., 2024; Chen et al., 2024c; Abdin et al., 2024; Liang et al., 2023a).

Despite their successes, the robustness of MLLMs to adversarial perturbations remains largely unexplored. This gap is critical as it challenges the reliability and safety of deploying MLLMs in real-world scenarios (Li et al., 2024c; Xie et al., 2024; Zhao et al., 2023; Zhang et al., 2024a). Adversarial perturbations–subtle changes to input data designed to deceive models–can significantly impact the performance and decision-making processes of these models, leading to harmful outcomes.

To bridge this gap, we introduce the RoMMath benchmark, which is designed to systematically assess the RObustness of MLLMs on Multimodal MATH reasoning against adversarial perturbations. RoMMath spans three primary areas: geometry, function, and statistic, ensuring the coverage of diverse and challenging scenarios. We construct an ***original*** **set** consisting of 600 multimodal math problems that span diverse mathematical and visual contexts. To systematically test the robustness of MLLMs at text- and vision-levels,

*Equal contributions. Corresponding author: Yilun Zhao (yilun.zhao@yale.edu).

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 11653–11665

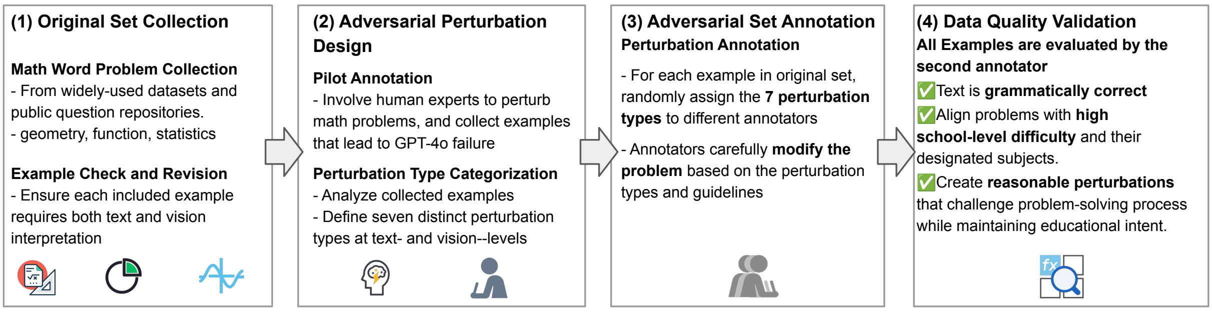

Figure 2: An overview of the ROMMATH benchmark construction pipeline.

we conduct a pilot study with expert annotators, identifying and designing seven types of adversarial perturbations to construct the ***adversarial* sets**. For each problem in the *original* set, expert annotators are assigned to develop multiple adversarial examples, each subjected to one type of perturbation. As a result, ROMMATH contains a total of 600 examples in the *original* set and 4,200 in the *adversarial* sets.

Figure 1 outlines the three research questions investigated in this study. We first conduct extensive experiments on ROMMATH, evaluating 17 MLLMs from 13 organizations known for their leading performance in multimodal math reasoning. Our experimental results reveal that current MLLMs generally exhibit performance drops when facing adversarial perturbations. For example, under the vision-level interference, the accuracy of the best-performing open-source model (*i.e.,* InternVL2.5 8B) falls from 55.7% to 46.3%. To gain deeper insights into the limitations of current MLLMs under adversarial perturbations, we conduct a thorough error analysis, categorizing five common error types these models exhibit. Finally, we investigate various strategies to improve model robustness, including detailed evaluations of in-context learning and prompting techniques.

Our contributions are summarized as follows:

- We introduce ROMMATH, a comprehensive benchmark designed to systematically evaluate the robustness of MLLMs in multimodal math reasoning when faced with adversarial perturbations. We design and annotate seven types of adversarial perturbations at text- and vision-levels, providing a systematic assessment (§2).
- We conduct a comprehensive evaluation of 17 MLLMs and reveal that current models generally exhibit significant performance drops when facing adversarial perturbations (§3).
- We conduct a thorough error analysis of both open-source and proprietary MLLMs with human experts, facilitating targeted improvement for future research (§4).
- We explore several strategies to improve the capabilities and robustness of MLLMs, offering valuable insights for future advancements (§5).

2 ROMMATH Benchmark

To provide a *systematic* and *diagnostic* evaluation of MLLM performance and robustness, ROMMATH adheres to the following data collection principles: **(1) Diverse Mathematical and Visual Contexts:** The benchmark should cover a wide range of mathematical and visual contexts. In response, ROMMATH spans three primary areas: *geometry*, *function*, and *statistics*, complemented by a variety of visual contexts (*e.g.,* diagrams, plots, and tables), to fully test the model's robustness in multimodal math reasoning (§2.1). **(2) Diagnostic Comprehensiveness:** The benchmark should provide various diagnostic angles on MLLM robustness. We design seven perturbations on text-level and vision-level for a systematic evaluation (§2.2). **(3) Reasonable Adversarial Perturbation:** The adversarial perturbation should be reasonable and meaningful, challenging the problem-solving process effectively. We ensure this through comprehensive human validation on each annotated example (§2.2).

In our preliminary study, we found it difficult to maintain data quality using annotators from Amazon Mechanical Turk. Therefore, we enlisted nine graduate students who are fluent in English and majoring in STEM fields for the dataset construction. We present an overview of the ROMMATH construction pipeline in Figure 2; and detail each construction process in the following subsections.

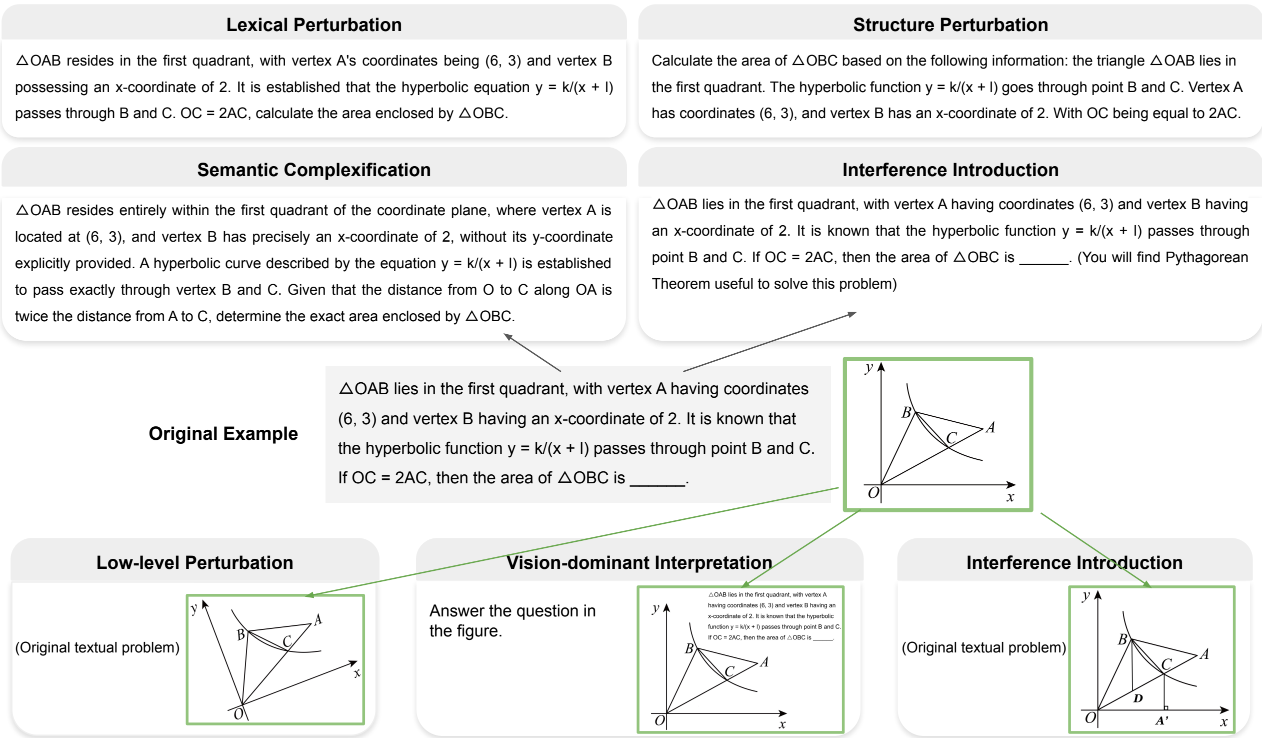

Figure 3: An example of the original problem (middle), with its corresponding text-level perturbations (top) and vision-level perturbations (bottom).

2.1 Original Set Collection and Annotation

We include three primary subjects – *geometry*, *function*, and *statistics* – into ROMMATH. ROMMATH focuses on high school-level math problems, ensuring they are challenging yet accessible to well-educated non-experts. By doing so, we avoid the complexity of advanced college-level mathematical topics like calculus and graph theory (Chen et al., 2023; Yue et al., 2024a). While some recent benchmarks (Lu et al., 2024b) include additional tasks like numeric commonsense QA and puzzle tests, we limit our scope to the three core subjects mentioned, as our focus is on foundational math reasoning capabilities.

We conduct a meticulous review, excluding problems that exhibit inappropriate difficulty or unsuitable formats. Additionally, the annotators must confirm that solving the math problem requires *both textual and visual information*. This process results in a total of 600 problems remaining within the ROMMATH *original* set.

2.2 Adversarial Perturbation Design

To ensure comprehensive coverage of perturbation types in this study, we begin with a **pilot annotation** phase involving multiple expert annotators. Specifically, we randomly select 30 examples that could be solved correctly by GPT-4o. We then engage five annotators and two of the authors to creatively perturb these questions, with the goal of introducing modifications that would cause GPT-4o to fail. We gather a total of 231 qualified examples. These examples are then reviewed by the core authors and categorized into seven distinct types that span text- and vision-level perturbations, as illustrated in Figure 3.

Text-level Perturbation:

ROMMATH includes the following four types of text-level perturbations:

(1) Lexical Perturbation Alter individual words or phrases within the text of a math problem without changing the overall structure. For example, replace words with less common synonyms or more specialized terminology.

(2) Structure Perturbation Modify the syntactic structure of the problem's textual statement. For example, change the order in which information is presented or rephrasing the statements using different grammatical structures.

(3) Semantic Complexification Enhance the problem's complexity by making its meaning more intricate. This can be done by introducing more complex relationships between elements within problems or by presenting problems in a more convoluted way. It aims to make the problem's text content more challenging to interpret.

(4) Interference Introduction Include misleading or distracting textual information that is irrelevant to the solution. For example, include extra numerical data, irrelevant background context, or misleading statements, requiring models to focus on pertinent information while ignoring the noise.

Vision-level Perturbation:

ROMMATH also includes the following three types of vision-level perturbations:

(1) Low-Level Perturbation Make subtle changes to the image in problem, *e.g.,* alter colors, brightness, contrast, or add minor visual noise.

(2) Vision-dominant Interpretation Move key information within the text to the image, requiring models to rely dominantly on the visual context to solve the problem.

(3) Interference Introduction Introduce visual elements that can distract or mislead the model. For example, include irrelevant or misleading components, extraneous symbols, or auxiliary lines that are not related to the original problem.

2.3 Adversarial Example Annotation

For each example in the *original* set, we randomly assign the seven perturbation types to different annotators. Each perturbation type corresponds to a unique adversarial version of the original example. This random assignment ensures that the perturbations cover a wide range of strategies, making the adversarial set diverse and comprehensive. Annotators are required to perturb the question according to the specific definitions and guidelines provided for their assigned perturbation type.

2.4 Data Quality Validation

To ensure the high quality of our ROMMATH dataset, particularly the *adversarial* sets, each annotated example is evaluated by another annotator based on the following criteria: (1) Text within the math problems should be grammatically correct and maintain clarity. (2) The problems should be appropriate for high school-level mathematics in terms of difficulty. (3) The adversarial perturbations should be reasonable and meaningful, challenging the problem-solving process effectively.

Property	**Testmini Set**	**Test Set**
Original Set (§2.1)		
Total Questions	200	400
Geometry	66	149
Function (new)	72	136
Statistics (new)	62	115
Multiple-choice Questions	138	270
Choices per Question	4	4
Free-form Questions	62	130
Question Length (Avg. / Max.)	52.5 / 190	51.8 / 188
Adversarial Set (§2.2)		
Total Perturbation Types (§ 2.2)	7	7
Text-level	4	4
Vision-level	3	3
Total Questions	200×8 $= 1{,}600$	400×8 $= 3{,}200$
Question Length (Avg. / Max.)	39.5 / 175	40.4 / 164
Total Examples	200+1,400 $= 1{,}600$	400+2,800 $= 3{,}200$

Table 1: Data Statistics of ROMMATH.

The validators are asked to revise examples that do not meet these standards. In practice, 451 out of 4,800 examples were revised by validators.

2.5 Data Statistics and Benchmark Release

Table 1 presents the key statistics of ROMMATH. We randomly divide the benchmark into two subsets: *testmini* and *test*. The *testmini* set is intended for model development validation, while It contains 200 original examples and 1,400 corresponding adversarial perturbations. the *test* set is designed for standard evaluation. It comprises the remaining 400 original examples and 2,800 corresponding perturbations. To prevent data contamination (Jacovi et al., 2023; Shi et al., 2024), the ground-truth answer for the *test* set will not be publicly released. Instead, we will develop and maintain an online evaluation platform, allowing researchers to evaluate models and participate in a public leaderboard.

3 Main Experiments

This section discusses the experimental setup and key findings from our main experiments, with a focus on addressing the first research question:

> 💡 **RQ1:** How robust are different MLLMs when performing multimodal math reasoning against different adversarial perturbations?

Model	Orig	Text-level Perturb				Vision-level Perturb			Avg
		Lexical	Structure	SemComp	Interfer.	Low-lvl.	V-dom.	Interfer.	
Human Expert	93.3	(	86.7		)	(	90.0	)	
Gemini-2-Flash	59.8	59.3	59.0	54.2	55.5	58.7	56.3	57.1	57.2
Grok-2-vision	55.0	56.0	53.8	51.3	50.3	54.7	40.7	47.2	50.6
GPT-4o	51.8	50.3	49.7	51.3	47.5	50.7	49.3	50.0	49.8
Claude-3.5-sonnet	49.7	52.3	48.5	49.5	49.5	50.5	47.8	48.3	49.5
InternVL2.5-8B	55.7	51.2	52.8	48.0	47.8	47.7	38.3	46.3	47.4
GPT-4o-mini	49.8	46.2	45.3	46.0	44.8	48.8	44.3	46.2	45.9
InternVL2-8B	50.5	43.2	46.2	41.2	44.7	45.8	34.8	40.5	42.3
LLaVA-onevision-7B	47.7	42.3	43.7	45.3	42.7	42.5	24.0	35.6	39.4
Qwen2-VL-7B	43.2	41.5	42.5	39.2	43.3	36.0	34.0	37.2	39.1
Pixtral-12b	36.5	38.0	35.2	37.7	37.7	37.7	26.8	35.1	35.5
Llama-3.2-V-11B	38.3	36.2	37.8	43.8	34.5	36.0	27.7	31.2	35.3
Idefics3-8B-Llama3	34.3	31.7	33.2	31.8	33.2	31.8	22.5	30.6	30.7
Molmo-7B-D	31.8	29.8	32.8	36.5	30.2	29.7	25.5	27.2	30.2
GLM-4V-9B	31.3	29.5	26.2	29.5	30.3	30.7	22.2	26.5	27.8
Phi-3	29.2	29.0	26.7	27.3	24.3	26.3	19.8	22.6	25.1
LLaVA-Next-8b	22.7	21.2	24.2	28.5	22.2	21.5	14.3	17.9	21.4
h2ovl-mississippi-2B	26.3	21.7	21.8	21.3	20.0	21.3	17.2	19.8	20.4

Table 2: Performance of MLLM on the ROMMATH *test* set. Average accuracy on the adversarial set is used as the ranking indicator. Cell colors indicate the rate of change in accuracy on the perturbation set compared to the original set, with red indicating a decrease, and green indicating an increase. "SemComp" refers to "Semantic Complexification", and "V-dom." refer to "Vision-dominant Interpretation", respectively.

3.1 Experiment Setup

Answer Accuracy Evaluation. Following previous work (Liang et al., 2023c; Zhang et al., 2024b; Lu et al., 2024b), we use **accuracy** as our evaluation metrics. Our evaluation pipeline adopts the approach used by MathVista (Lu et al., 2024b), which involves applying GPT-4o to extract answer text, normalizing this text to the required answer format (*e.g.*, an option letter for multi-choice questions), and then computing the accuracy scores.

Baseline and Human-level Performance. We also set up several baselines for performance comparison: (1) *random chance*, where we select one option at random for multiple-choice questions and leave free-form questions blank; and (2) *frequent chance*, where we choose the most frequent answer for multiple-choice and free-form questions, separately. We also measure the *Human Expert Performance* on ROMMATH. Specifically, we enlisted four evaluators and randomly distributed 120 different examples among them. These 120 examples were composed of 40 sets of problems. Each set included one sample from the original set and its corresponding samples from the text-level and vision-level adversarial sets. To prevent leakage effects caused by evaluators completing problems with the same original source, we randomly assigned each example to the evaluators without providing any hints about the perturbations and ensured that each evaluator only completed one problem from each set.

Evaluated MLLMs. We examine the performance of 17 MLLMs across two distinct categories on ROMMATH: (1) **Open-source MLLMs**, including LLaVA (Liu et al., 2023, 2024), Qwen2-VL and Qwen2.5-VL (Wang et al., 2024b), GLM-4V (GLM et al., 2024), Molmo (Deitke et al., 2024), Pixtral-12B (Dong et al., 2024), H2OVL (Galib et al., 2024), Idefics3 (Laurençon et al., 2024), Phi-3.5-Vision (Abdin et al., 2024), Llama-3.2-Vision (Meta, 2024), InternVL2 (Ailab, 2024), and InternVL2.5 (Chen et al., 2025). (2) **Proprietary MLLMs**, including GPT-4o & GPT-4o-mini (OpenAI, 2024), Grok-2-Vision (xAI, 2024), Claude-3.5 (Anthropic, 2024b), and Gemini-2.0-flash (Gemini, 2024). Appendix A presents the details of evaluated models. Following previous work on multimodal reasoning (Yue et al., 2024a; Lu et al., 2024b), by default, our experiments are conducted under *zero-shot Chain-of-Thought* settings to assess the generalization capacity of MLLMs without few-shot prompting or further fine-tuning. The employed CoT prompt is presented in Table 3.

3.2 Experimental Results

Table 2 presents the MLLM performance on ROMMATH. We first analyze the results on ***original* set** and summarize our key findings as follows:

ROMMATH presents substantial challenges for current MLLMs. A significant performance gap is observed between human experts and evaluated MLLMs on the *original* set of ROMMATH. Notably, Gemini-2-Flash, the highest-performing model to date, achieves an accuracy rate of only 59.8%, in contrast to the 93.3% accuracy of human experts. Moreover, the evaluated proprietary MLLMs generally exhibit better performance than open-source MLLMs. These discrepancies highlight the complexity and challenges of the original problems collected in our benchmark. We believe that the ROMMATH *original* set is valuable on its own in assessing MLLM performance.

We then analyze the MLLM performance against adversarial perturbations and present the following key findings, with more detailed analyses provided in the subsequent sections.

Human performance maintains consistency on the original and adversarial sets. The results reveal that human experts are largely unaffected by the annotated adversarial perturbations, consistent with the "Reasonable Adversarial Perturbation" data collection principle outlined in Section 2. This demonstrates that the adversarial perturbations in ROMMATH are reasonable and should not hinder those individuals with strong and robust reasoning abilities.

The critical challenge of MLLM robustness against adversarial perturbations needs greater attention from the research community. Both open-source and proprietary MLLMs generally exhibit significant performance drops when facing adversarial perturbations. This highlights the vulnerability of current MLLMs to adversarial perturbations, underscoring the need for improved model robustness in future work. Among various text- and vision-level perturbation types, *interference introduction* causes the significant performance degradation. This highlights the weakness of current MLLMs in distinguishing key information from irrelevant and misleading noise, especially when visual interpretation is required. However, we observe a notable performance improvement in open-source models released recently. In particular, the InternVL2.5 achieves performance comparable to proprietary models, underscoring the potential of open-source models through continued innovation and community collaboration.

4 MLLM Error Analysis

> 💡 **RQ2:** On ROMMATH, what common errors can be identified in MLLM reasoning, especially under adversarial perturbations?

To answer RQ2, we conduct an in-depth human analysis of the error cases made by GPT-4o, Qwen2-VL-7B, and Llama 3.2-Vision 11B, as they achieve substantial performance among proprietary and open-source MLLMs. Specifically, for each model, we randomly sample (1) 50 error examples from the *testmini original* set and (2) 50 examples that are correctly solved in the *testmini original* set but fail under adversarial perturbations. Through in-depth analysis, we have identified the following five common error types that current MLLMs are likely to make:

(1) Reasoning Error: The model lacks or incorrectly applys logical reasoning to solve the problem, such as missing critical steps or making improper reasoning.

(2) Vision Misinterpretation: The model makes errors when interpreting or extracting elements and their attributes from charts, such as numbers, geometric shapes, element matching, and annotation relationships.

(3) Text Misinterpretation: The model misunderstands the given conditions of the textual problem. It may involve misreading or misinterpreting key terms, instructions, or numerical values within the problem statement.

(4) Text and Vision Misalignment: The model mismatches the information presented in the image and the question, such as misplaced objects or skewed angles, hinders accurate correlation and understanding.

(5) Calculation Error: The model makes errors in mathematical calculation, leading to incorrect numerical results.

5 Exploring Strategies to Enhance MLLM Robustness

Building on the analysis from RQ2, we explore several potential strategies for enhancing model ro-

Prompt Variants	Prompt
Standard CoT (used for main experiments)	Solve the math world problem using the provided textual and visual context. You should first conduct reasoning step by step, and then provide the final answer at the end.
Direct Output	Solve the math world problem using the provided textual and visual context. You should directly output the final answer without providing reasoning process.
Descibe-then-Reason CoT	Solve the math word problem using the provided textual and visual context. You should conduct reasoning step by step in the format [Problem Overview, Step-by-Step Detailed Solution, Final Answer], 'Problem Overview' should contain the information you learn from the text and image respectively.

Table 3: Variants of zero-shot prompting methods investigated in RQ3.

bustness on the RoMMath *testmini* set. For these experiments, we utilize the two top-performing *open-source* models, Llama 3.2-Vision 11B and Qwen2-VL-7B, to gain insights addressing RQ3:

RQ3: What strategies can be explored in future research to enhance the performance and robustness of MLLMs against adversarial perturbations?

5.1 In-Context Learning

As discussed in Section 3.1, our main experiments are conducted in zero-shot settings to assess the *generalization capacity* of MLLM without example demonstrations. However, we believe that utilizing few-shot settings with example demonstrations could potentially enhance model performance and robustness. To test this hypothesis, we compare several variants in a one-shot setting using examples, along with human-annotated step-by-step solutions, from different sources: (1) from the *original* set; (2) from the *adversarial* sets and with a different perturbation type; and (3) from the *adversarial* sets and with the same type of perturbation as the tested one. To ensure fairness, we employ the same standard CoT prompting for experiments. We randomly sample one math problem from the *testmini* set for the example demonstration. As illustrated in Table 5, providing example demonstrations generally improves MLLM performance on both original and adversarial sets. Additionally, using examples from the *adversarial* sets to demonstrate how to handle adversarial perturbations can help reduce the performance gap caused by the adversarial perturbations.

5.2 Different Prompting Methods

We also investigate the impact of different instructions within zero-shot prompts on the model's performance and robustness. We design three variants of zero-shot prompts (presented in Table 3), specifically: (1) *Direct Output*: the MLLM is instructed to directly output the final answer without performing step-by-step reasoning. (2) *Standard CoT*: the same as used in the main experiments, where the model is instructed to perform step-by-step reasoning before providing the final answer. (3) *Describe-then-Reason CoT* (Jia et al., 2024): the model is instructed to first re-describe the math

Systems	Orig. Set	Adv. Set
Llama 3.2-Vision 11B		
0-shot CoT	38.3	35.3
1-shot CoT		
From original set	38.6 (+0.3)	35.8 (+0.5)
From different type	39.1 (+0.8)	36.5 (+1.2)
From same type	39.0 (+0.7)	37.2 (+1.9)
Qwen2-VL-7B		
0-shot CoT	43.2	39.1
1-shot CoT		
From original set	44.0 (+0.8)	40.1 (+1.0)
From different type	44.3 (+1.1)	40.3 (+1.2)
From same type	43.8 (+0.6)	40.6 (+1.5)

Table 4: Result analyses of Llama 3.2-Vision 11B and Qwen2-VL-7B on the *testmini* set under different in-context learning setting.

Systems	Orig. Set	Adv. Set
Llama 3.2-Vision 11B		
Standard CoT	38.3	35.3
Direct Output	36.9 (-1.4)	33.0 (-2.3)
Describe-then-Reason CoT	40.6 (+2.3)	38.5 (+3.2)
Qwen2-VL-7B		
Standard CoT	43.2	39.1
Direct Output	41.6 (-1.6)	37.0 (-2.1)
Describe-then-Reason CoT	44.2 (+1.0)	40.5 (+1.4)

Table 5: Result analyses of Llama 3.2-Vision 11B and Qwen2-VL-7B on the *testmini* set using different prompting methods.

problem, interpreting and extracting all key textual and visual information before proceeding with step-by-step reasoning. This aligns with our findings in RQ2 that more reasoning errors occur in the step of interpreting and extracting information from the textual and visual context. Therefore, we adopt this idea in our benchmark. As illustrated in Table 5, Describe-then-Reason CoT achieves the best performance and robustness.

6 Related Work

Multimodal Math Reasoning Benchmark. With the growing interest in evaluating the reasoning capabilities of foundation models across textual and visual contexts, multimodal math reasoning benchmarks have gained prominence. Early benchmarks primarily focused on specific areas such as geometry (Seo et al., 2015; Lu et al., 2021; Chen et al., 2022; Cao and Xiao, 2022) and chart interpretation (Kahou et al., 2017; Methani et al., 2020; Masry et al., 2022b; Wang et al., 2024c). Recently, comprehensive datasets (Lu et al., 2024b; Wang et al., 2024a; Chen et al., 2024a; Sun et al., 2024; Yue et al., 2024b) have been developed to cover a broad spectrum of specialized multimodal mathematical tasks. However, despite these advancements, the robustness of MLLMs in multimodal math reasoning remains under-explored.

Evaluating Robustness of Foundation Model. Evaluating robustness of multimodal foundation models has become a critical area of research. Benchmarks like Avibench (Zhang et al., 2024a) and ChartInsights (Wu et al., 2024) primarily target low-level image perturbations, examining a variety of distortions such as noise, blur, weather effects, font size, and image element transformations. Additionally, initiatives like MathVerse (Zhang et al., 2024b) and MMStar (Chen et al., 2024a) emphasize balancing models' visual and textual capabilities. However, these methods are mainly centered on low-level perturbations and do not thoroughly analyze how such image changes induce specific errors or affect the models' reasoning abilities. In contrast, our research shifts the focus towards understanding the impacts of adversarial perturbations on model reasoning. We develop ROMMATH and conduct a systematic evaluation of MLLM robustness against adversarial perturbations.

7 Conclusion

This paper presents ROMMATH, a new benchmark designed to evaluate the robustness of MLLMs in multimodal math reasoning against adversarial perturbations. We reveal a significant decline in MLLMs performance under adversarial conditions. Through detailed error analysis by human experts, we gain a deeper understanding of the current limitations of MLLMs. We also explore various strategies, including in-context learning and different prompting methods, to enhance both model performance and robustness, providing insights for future research.

Limitations and Future Work

In this work, we perform a comprehensive analysis of MLLMs' capabilities and robustness on multimodal math reasoning tasks. However, our work still has some limitations: First, recent works (Zhao et al., 2023; Sheshadri et al., 2024) have shown that training foundation models on adversarial data can enhance their robustness. However, due to computational constraints, we do not explore the adversarial training in our study. Instead, our study does not explore adversarial training. Instead, we focus on improving model robustness through in-context learning and advanced prompting techniques. We encourage future research to investigate the application of model training methods (Liang et al., 2023b; Sheshadri et al., 2024; Liang et al., 2024b) on ROMMATH for further robustness improvements. Moreover, this study focuses on high-school-level multimodal math reasoning and does not extend to more advanced topics such as calculus or graph theory (Chen et al., 2024b; He et al., 2024). The primary objective is to examine the robustness of models in multimodal mathematical reasoning when faced with adversarial perturbations. High-school-level problems are chosen as they present a manageable level of complexity, enabling us to focus on robustness without the additional challenges posed by more advanced reasoning and domain-specific knowledge. We believe that future work, as multimodal language models continue to evolve, could extend our work by evaluating model robustness in more sophisticated mathematical reasoning tasks.

Acknowledgments

We are grateful to Google TRC program for providing computing resources and Together AI for granting LLM API credits for this project.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

Ailab. 2024. Internvl2: Better than the bestexpanding performance boundaries of open-source multimodal models with the progressive scaling strategy.

Anthropic. 2024a. Introducing claude 3.5 sonnet.

Anthropic. 2024b. Introducing the next generation of claude.

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 Xing, and Liang Lin. 2021. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523, Online. Association for Computational Linguistics.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024a. Are we on the right way for evaluating large vision-language models?

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. TheoremQA: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901, Singapore. Association for Computational Linguistics.

Xiuying Chen, Tairan Wang, Taicheng Guo, Kehan Guo, Juexiao Zhou, Haoyang Li, Mingchen Zhuge, Jürgen Schmidhuber, Xin Gao, and Xiangliang Zhang. 2024b. Scholarchemqa: Unveiling the power of language models in chemical research question answering.

Zhe Chen, Weiyun Wang, Yue Cao, Zhangwei Gao, Erfei Cui, Jingou Zhu, and et al. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv*, 2412.05271.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Jen Dumas, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. https://www.allenai.org/.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Shaikat Galib, Shanshan Wang, Guanshuo Xu, Pascal Pfeiffer, Ryan Chesler, Mark Landry, and Sri Satish Ambati. 2024. H2ovl-mississippi vision language models technical report. Technical report, H2O.ai. https://www.h2o.ai/.

Gemini. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand. Association for Computational Linguistics.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore. Association for Computational Linguistics.

Mengzhao Jia, Zhihan Zhang, Wenhao Yu, Fangkai Jiao, and Meng Jiang. 2024. Describe-then-reason: Improving multimodal mathematical reasoning through visual comprehension training.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024a. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024b. LLaVA-OneVision: Easy Visual Task Transfer. https://arxiv.org/abs/2408.03326.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. 2024c. Gsm-plus: A comprehensive benchmark for evaluating the robustness of llms as mathematical problem solvers.

Zhenwen Liang, Kehan Guo, Gang Liu, Taicheng Guo, Yujun Zhou, Tianyu Yang, Jiajun Jiao, Renjie Pi, Jipeng Zhang, and Xiangliang Zhang. 2024a. SceMQA: A scientific college entrance level multimodal question answering benchmark. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 109–119, Bangkok, Thailand. Association for Computational Linguistics.

Zhenwen Liang, Tianyu Yang, Jipeng Zhang, and Xiangliang Zhang. 2023a. UniMath: A foundational and multimodal mathematical reasoner. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7126–7133, Singapore. Association for Computational Linguistics.

Zhenwen Liang, Dian Yu, Xiaoman Pan, Wenlin Yao, Qingkai Zeng, Xiangliang Zhang, and Dong Yu. 2024b. MinT: Boosting generalization in mathematical reasoning via multi-view fine-tuning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11307–11318, Torino, Italia. ELRA and ICCL.

Zhenwen Liang, Wenhao Yu, Tanmay Rajpurohit, Peter Clark, Xiangliang Zhang, and Ashwin Kalyan. 2023b. Let GPT be a math tutor: Teaching math word problem solvers with customized exercise generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14384–14396, Singapore. Association for Computational Linguistics.

Zhenwen Liang, Jipeng Zhang, Kehan Guo, Xiaodong Wu, Jie Shao, and Xiangliang Zhang. 2023c. Compositional mathematical encoding for math word problems. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10008–10017, Toronto, Canada. Association for Computational Linguistics.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024a. Deepseek-vl: Towards real-world vision-language understanding.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024b. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

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. *arXiv preprint arXiv:2105.04165*.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022a. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022b. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Meta. 2024. meta-llama/llama-3.2-11b-vision-instruct.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536.

OpenAI. 2024. Hello gpt-4o.

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.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. 2024. Targeted latent adversarial training improves robustness to persistent harmful behaviors in llms.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*.

Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juanzi Li. 2024. Mm-math: Advancing multimodal math evaluation with process evaluation and fine-grained classification.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024c. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *ArXiv*, abs/2406.18521.

Yifan Wu, Lutao Yan, Yuyu Luo, Yunhai Wang, and Nan Tang. 2024. Evaluating task-based effectiveness of mllms on charts.

xAI. 2024. Grok-2 beta release.

Roy Xie, Chengxuan Huang, Junlin Wang, and Bhuwan Dhingra. 2024. Adversarial math word problem generation.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, et al. 2024b. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Hao Zhang, Wenqi Shao, Hong Liu, Yongqiang Ma, Ping Luo, Yu Qiao, and Kaipeng Zhang. 2024a. Avibench: Towards evaluating the robustness of large vision-language model on adversarial visual-instructions.

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

Yilun Zhao, Chen Zhao, Linyong Nan, Zhenting Qi, Wenlin Zhang, Xiangru Tang, Boyu Mi, and Dragomir Radev. 2023. RobuT: A systematic study of table QA robustness against human-annotated adversarial perturbations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6064–6081, Toronto, Canada. Association for Computational Linguistics.

A Appendix

Model Series	Organization	Release	Source
GPT-4o (OpenAI, 2024)	OpenAI	2024-11	https://platform.openai.com/docs/models/gpt-4o
Claude-3.5-sonnet (Anthropic, 2024a)	Anthropic	2024-10	https://www.anthropic.com/api
Gemini-2.0-flash (Gemini, 2024)	Google	2024-11	https://ai.google.dev/gemini-api/docs
Grok-2-Vision (xAI, 2024)	xAI	2024-8	https://docs.x.ai/docs/models?cluster=us-east-1
Qwen2-VL (Wang et al., 2024b) Qwen2.5-VL (Wang et al., 2024b)	Qwen Team	2024-9 2025-1	Qwen/Qwen2-VL-*B-Instruct Qwen/Qwen2.5-VL-7B-Instruct
Idefics3 (Laurençon et al., 2024)	Hugging Face	2024-8	HuggingFaceM4/Idefics3-8B-Llama3
LLaVA-NeXT (Li et al., 2024a) LLaVA-Onevision(Li et al., 2024b)	LMMs-lab	2024-4 2024-8	lmms-lab/llama3-llava-next-8b-hf lmms-lab/llava-onevision-qwen2-7b-ov
Phi-3.5-Vision (Abdin et al., 2024)	Microsoft	2024-7	microsoft/Phi-3.5-vision-instruct
H2OVL (Galib et al., 2024)	H2O AI	2024-10	h2oai/h2ovl-mississippi-2b
GLM-4V (GLM et al., 2024)	THUDM	2024-8	THUDM/glm-4v-9b
Molmo (Deitke et al., 2024)	Allen Institute for AI	2024-9	allenai/Molmo-7B-D-0924
Pixtral-12B (Dong et al., 2024)	Mistral AI	2024-9	mistralai/Pixtral-12B-2409
Llama-3.2-Vision (Meta, 2024)	Meta	2024-9	meta-llama/Llama-3.2-11B-Vision-Instruct
InternVL2 (Ailab, 2024) InternVL2.5 (Chen et al., 2025)	Shanghai AI Lab	2024-7 2025-2	OpenGVLab/InternVL2-*B OpenGVLab/InternVL2.5-*B

Table 6: Details of the organization, release time, and model source (*i.e.,* url for proprietary models and Huggingface model name for open-source models) for the LLMs evaluated in ROMMATH.