MPBench: A Comprehensive Multimodal Reasoning Benchmark for Process Errors Identification

Zhaopan Xu[1,2], **Pengfei Zhou**[3], **Jiaxin Ai**[4,5], **Wangbo Zhao**[3], **Kai Wang**[3], **Xiaojiang Peng**[6]
Wenqi Shao[2,5], **Hongxun Yao**[1]*, **Kaipeng Zhang**[2,5]*
[1]Harbin Institute of Technology, [2]Shanghai AI Laboratory,[3]NUS,
[4]Wuhan University,[5]Shanghai Innovation Institude,[6]SZTU
`h.yao@hit.edu.cn, zhangkaipeng@pjlab.org.cn`
`https://mpbench.github.io`

Abstract

Reasoning is an essential capacity for large language models (LLMs) to address complex tasks, where the identification of process errors is vital for improving this ability. Recently, process-level reward models (PRMs) were proposed to provide step-wise rewards that facilitate reinforcement learning and data production during training and guide LLMs toward correct steps during inference, thereby improving reasoning accuracy. However, existing benchmarks of PRMs are text-based and focus on error detection, neglecting other scenarios like reasoning search. To address this gap, we introduce MPBench, a comprehensive, multi-task, multimodal benchmark designed to systematically assess the effectiveness of PRMs in diverse scenarios. MPBench employs three evaluation paradigms, each targeting a specific role of PRMs in the reasoning process: (1) Step Correctness, which assesses the correctness of each intermediate reasoning step; (2) Answer Aggregation, which aggregates multiple solutions and selects the best one; and (3) Reasoning Process Search, which guides the search for optimal reasoning steps during inference. Through these paradigms, MPBench makes comprehensive evaluations and provides insights into the development of multimodal PRMs.

1 Introduction

Artificial intelligence (AI) has not yet effectively addressed complex reasoning tasks such as mathematics, programming, and planning. A barrier to achieving artificial general intelligence (AGI) remains this. To address this challenge, the recent release of the GPT-o1 (OpenAI, 2024) proposed learning to reason to master human-like reasoning processes. Unlike traditional LLMs' thinking, reasoning models generate long chains of thought. Consequently, recent research has increasingly paid attention to process-level (i.e., stepwise) reasoning analysis and further introduced various reasoning models to achieve significant enhancements in tasks such as mathematics and code generation.

Identifying process errors can facilitate effective reasoning trajectories. PRMs, introduced by (Uesato et al., 2022) and (Lightman et al., 2023), can provide stepwise feedback for multi-step reasoning results. By rewarding intermediate steps, PRMs can enhance LLMs' reasoning capabilities during both training and inference. Specifically, PRMs facilitate reinforcement learning or chain-of-thought data generation by offering fine-grained rewards, thus reducing the dependence on human-annotated data and improving performance. Moreover, during inference, PRMs guide LLMs in evaluating and exploring intermediate "thoughts," promoting the generation of more deliberate reasoning steps and ultimately leading to improved reasoning accuracy.

Despite the critical role of PRMs, they have not been adequately evaluated. Best-of-N performance, a widely employed evaluation paradigm, is time-consuming, lacks finer-grained inspection, and its evaluation reliability can be significantly affected by the underlying solution generation model. This limited scope of evaluation hinders a comprehensive understanding of PRMs' potential to enhance complex reasoning tasks. Recent benchmarks (Zheng et al., 2024; Song et al., 2025; Ai et al., 2025) were introduced to evaluate PRMs. However, they typically focus on a single scenario, such as step-by-step error identification, evaluating the ability of PRMs to detect erroneous steps accurately. These evaluation paradigms inadequately assess PRM performance in some scenarios. For example, during the reasoning process search in LLM inference, PRMs are tasked with selecting the correct step from candidates without access to the complete reasoning trajectory and the final answer. Furthermore, whereas current PRM benchmarks are text-centric, multimodal benchmarks are

---

* Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 21586–21606
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

	PRM Benchmarks?	Multimodal Benchmarks?	Evaluation Paradigms	Step Annotation	Annotator	Test Case Size
MR-GSM8K (Zeng et al., 2023)	✗	✗	1	✓	Human	2,999
CriticBench (Lin et al., 2024)	✗	✗	1	✗	-	-
MathCheck-GSM (Zhou et al., 2024)	✗	✗	1	✓	Synthetic	516
M^3CoT (Chen et al., 2024a)	✗	✗	1	✓	Human	5,975
ProcessBench (Zheng et al., 2024)	✓	✗	1	✓	Human	3,400
PRMBench (Song et al., 2025)	✓	✗	1	✓	Synthetic + Human	6,216
MPBench (Ours)	✓	✓	3	✓	Synthetic + Human	9,745

Table 1: Comparison between reasoning-related LLM benchmarks with our MPBench benchmark.

deficient, while multimodal contents are common in real-world tasks.

These challenges motivate the development of **MPBench**, a comprehensive multimodal benchmark designed to evaluate the efficacy of multimodal Process-level Reward Models (PRMs) across three distinct scenarios. MPBench centers on three key evaluation paradigms: **Step Correctness**, which assesses PRMs' ability to evaluate the correctness of each intermediate reasoning step and provide stepwise rewards to support reinforcement learning; **Answer Aggregation**, where PRMs aggregate per-step scores from multiple candidate solutions to select the optimal one; and **Reasoning Process Search**, which examines PRMs' capacity to guide the search for optimal reasoning steps by enabling a structured exploration of potential solutions. Featuring 9,745 fine-grained data instances across these three scenarios, MPBench offers a robust framework for comprehensively assessing PRM performance in real-world reasoning tasks, thereby providing valuable insights into their role in enhancing the reasoning capabilities of MLLMs.

We conduct extensive experiments on MPBench with 14 MLLMs (prompted as critic models), including closed-source models, GPT-4o, Claude-3.7 and Gemini-2.0, and open-source models like InternVL (Chen et al., 2024b), QWenVL (Team, 2025), QVQ (Team, 2024), and VisualPRM (Wang et al., 2025). They struggle to achieve satisfactory results, and we believe our benchmarks benefit the future development of multimodal PRMs and process-level analysis. Our key contributions are summarized as follows:

- We present MPBench, the first comprehensive multimodal process-level reward model benchmark, comprising 9,745 fine-grained data instances across diverse subjects, tasks, and challenges.
- MPBench incorporates three distinct evaluation paradigms that comprehensively assess the role of PRMs in enhancing MLLM reasoning during both training and inference. These paradigms include step correctness, answer aggregation, and reasoning process search.
- We in-depth analyze the performance of 12 MLLMs and reveal distinct performance characteristics across different scenarios, providing valuable insights to assist future research on the development of multimodal PRMs.

2 Related Work

2.1 Reasoning Benchmarks

GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) have served as prominent benchmarks for evaluating the mathematical reasoning capabilities of LLMs, primarily focusing on assessing the final correctness of generated solutions. Subsequent work has explored synthesizing solutions and evaluating intermediate steps. For instance, MathCheck (Zhou et al., 2024) focuses on judging the correctness of individual reasoning steps. CriticBench (Lin et al., 2024) evaluates language models' abilities to critique solutions and extends evaluation to various reasoning tasks. With the growing interest in the development of process reward models (PRMs) to enhance model reasoning capabilities, several benchmarks are specifically designed for PRMs. such as ProcessBench (Zheng et al., 2024) and PRMBench (Song et al., 2025).

These reasoning benchmarks are primarily text-based. With the increasing demand for evaluating multimodal reasoning, there has been a drive to develop multimodal reasoning benchmarks across diverse domains (Zeng et al., 2024; Xu et al., 2025). However, research on benchmarks specifically tailored to multimodal PRMs remains limited. To address this gap, we introduce MPBench to benchmark multimodal process reward models. As highlighted in Table 1, MPBench encompasses a wide range of tasks, along with a large-scale, diverse set of tasks, enabling a more comprehensive evaluation of multimodal PRMs' reasoning capabilities.

2.2 Process Reward Models

The increasing prevalence of Process Reward Models (PRMs) highlights their crucial role in enhancing reasoning by providing step-wise rewards that facilitate reinforcement learning and data generation. To evaluate the accuracy of these step-wise rewards, benchmarks like ProcessBench and PRMBench have been developed. These benchmarks focus on constructing reasoning processes with erroneous steps for mathematical problems and evaluating the PRMs' ability to identify the first error or the fine-grained error category of each step. Recently, GPT-o1 has shown impressive reasoning trajectories, highlighting the potential of scaling test-time computation to enhance LLM reasoning accuracy (Thawakar et al., 2025; Zhao et al., 2024a,b). Building on this idea, several studies (Snell et al., 2024; Xiang et al., 2024) have explored using PRMs to guide search within the interactive process during inference, further boosting model reasoning capabilities. However, evaluating PRMs in this application scenario remains limited. This is primarily due to the need for group and tree-structured reasoning process data, which existing PRM benchmarks do not adequately address. This gap in evaluation is precisely one of the aspects our work intends to address.

3 MPBench

MPBench is a comprehensive benchmark designed to evaluate how Process-level Reward Models (PRMs) enhance reasoning, comprising 9,745 fine-grained instances. It employs three core evaluation paradigms: **Step Correctness**, which assesses PRMs' ability to judge the correctness of individual intermediate reasoning steps, providing stepwise rewards for reinforcement learning (Xia et al., 2024) and reducing reliance on human-annotated data; **Answer Aggregation**, which tasks PRMs with aggregating per-step scores from multiple candidate solutions to identify the best one; and **Reasoning Process Search**, which evaluates PRMs' capacity to guide the search for optimal reasoning steps during inference by enabling a tree search over the solution space, thereby encouraging more deliberate reasoning and improving accuracy.

3.1 Data Curation

Our dataset is based on M^3CoT (Chen et al., 2024a), a large-scale multimodal dataset comprising 17 topics and 263 categories across three primary domains: science knowledge, mathematics, and commonsense. M^3CoT provides questions, ground truth answers, and ground truth step-level solution processes. Following CoMT (Cheng et al., 2024), we filtered out low-quality instances (e.g., vague expressions) to establish our ground truth answers, ensuring the reliability and accuracy of our dataset.

We curated the dataset by extracting metadata and constructing test cases according to our defined evaluation paradigms. Corresponding to these paradigms, MPBench comprises three distinct data categories: (1) **Erroneous Steps:** This category evaluates PRM capacity to provide step-level supervision by identifying incorrect steps within solution sequences. (2) **Multi-solution:** This component tests PRM ability to rerank candidate answers during inference, selecting the most promising solution. (3) **Action Trees:** This structured data assesses PRM guidance during reasoning process search, where PRMs navigate a tree of possible actions to identify the optimal reasoning path.

Visual examples of data for these three evaluation paradigms can be found in Appendix Figs 6, 7, and 8. Comprehensive statistics for MPBench are presented in Appendix A. The detailed construction of erroneous steps, multi-solution, and action trees is provided below:

Erroneous Steps To generate erroneous steps, we leveraged GPT-4o to introduce reasonable errors into the ground truth reasoning processes.

Multi-solution To generate a diverse set of candidate solutions, including incorrect ones, we employed four readily available multimodal language models: two open-source (LLaVa and QWen) and two closed-source (GPT-4o and Gemini). Each model generated three solutions per problem. From these generated solutions, we randomly selected one solution from each model, along with the original ground truth solution, to create a final set of five candidate solutions. For problems where all generated solutions were correct, we randomly selected incorrect solutions from the pool of all generated incorrect solutions for that problem.

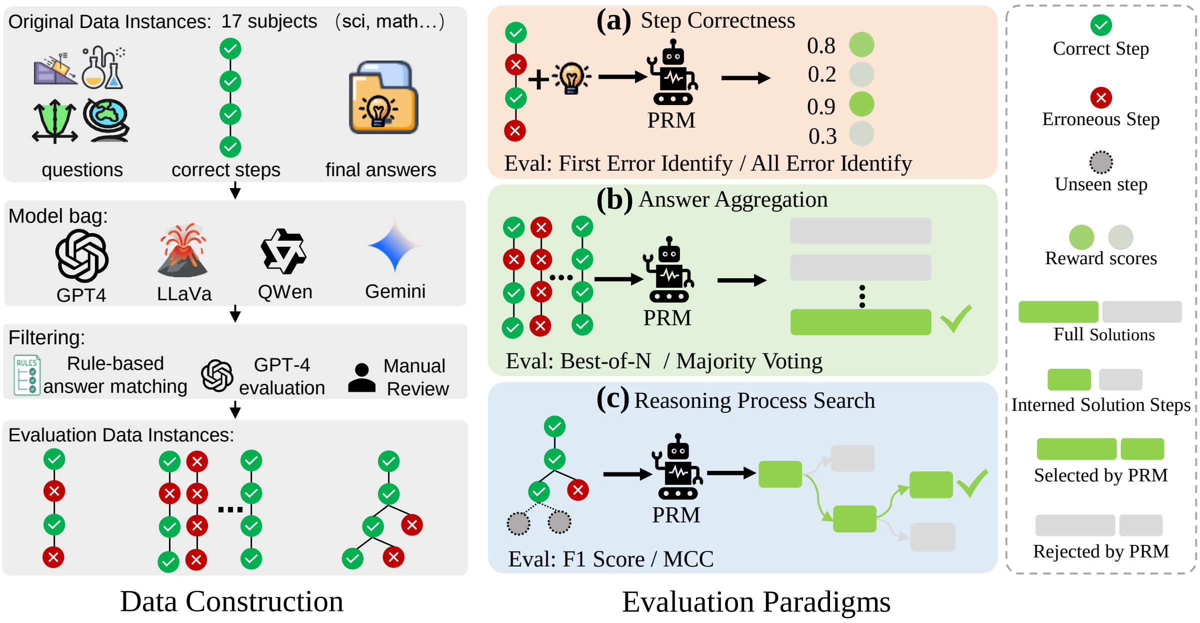


Figure 1: An overview of our MPBench. Left: data curation procedure. Right: evaluation paradigms: Step Correctness, Answer Aggregation, and Reasoning Process Search, highlighting the assessment of PRM performance through various tasks such as identifying errors, aggregating answers, and guiding reasoning steps

Action Trees To construct action trees, we prompted GPT-4o to expand each incorrect action into the corresponding ground truth steps. This expansion process yielded multiple action pairs for the Reasoning Process Search evaluation.

3.2 Evaluation Objectives

3.2.1 Step Correctness

This assessment takes annotated question-solution pairs and corresponding ground truth answers as input. The task is to evaluate the correctness of each individual step, generating step-level scores. A predefined threshold is then applied to these scores, yielding a binary prediction of step correctness. Consequently, this evaluation is framed as a binary classification problem, with the F1 score serving as the evaluation metric. Following ProcessBench (Zheng et al., 2024) and PRMBench (Song et al., 2025), which uses the negative F1 score as a metric for error detection, we also combine the negative F1 score ($F1_{neg}$) with the F1 score to compute a comprehensive evaluation metric called RMScore.

$$RM\text{-}Score = w_1 * F1_{neg} + w_2 * F1, \quad (1)$$

where w_1 and w_2 are weights, both set to 0.5 by default. Based on diverse application scenarios, we further categorize this assessment into two subcategories: First Error Identification and All Error Identification, as detailed below:

First Error Identification requires PRMs to identify the first error encountered in a reasoning process. This evaluation method is commonly employed in both PRM testing (Zheng et al., 2024) and training (Hwang et al., 2024).

All Error Identification This sub-category evaluates the PRMs' ability to identify all errors within a given solution. This comprehensive error identification is crucial for providing fine-grained rewards during training, enabling reinforcement learning.

3.2.2 Answer Aggregation

In this scenario, the Process-level Reward Model (PRM) is presented with a question and a diverse set of multiple candidate solutions. The primary objective for the PRM is to meticulously evaluate these candidates and ultimately identify the correct solutions among them. To assess the efficacy of PRMs in this critical task, we evaluate their performance across two distinct search approaches: Best-of-N and Majority Voting.

Best-of-N applies PRMs to score each candidate solution independently and selects the solution with the highest individual score as the final answer. It assumes that the highly-rated solution is indicative of overall correctness, making it a straightforward yet powerful mechanism for solution selection.

Majority Voting operates on the principle of collective consensus among inference traces. This approach focuses on the answers supported by the majority of the model's generated reasoning paths. The scores of responses associated with the same answer are aggregated, and the answer with the highest aggregated score is chosen as the final answer. This method aims to enhance robustness by mitigating the impact of individual outlier scores and leveraging the collective wisdom derived from multiple reasoning attempts.

3.2.3 Reasoning Process Search

This evaluation assesses search performance during inference. The PRM receives as input the question, the corresponding history of reasoning steps, and candidate steps for the current step under evaluation. In our dataset, this component consists of paired tree-structured data, where each step presents a binary choice. We initially use the F1 score as the primary evaluation metric. Furthermore, following the work of MR-Score (Zeng et al., 2024), we employ the Matthews Correlation Coefficient (MCC) (Matthews, 1975) for the binary classification of search correctness.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \quad (2)$$

where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative. The MCC score ranges from -1 to +1 with -1 meaning total disagreement between prediction and observation, 0 indicating near random performance, and +1 representing perfect prediction.

3.3 Quality Control

3.3.1 Filtering

To ensure high data quality and validity, and to maintain a reasonable level of challenge, we implemented a multi-stage filtering process, incorporating rule-based filtering, GPT-4 review, and a simple problem filter. Specifically, this process included: 1) defining rules to ensure adherence to the required format; 2) manually developing in-context (IC) examples to guide GPT-4 in identifying and filtering unreasonable instances; and 3) using Gemini for testing and filtering solutions where the absolute difference between the scores of incorrect and correct steps was greater than 1, thereby ensuring dataset difficulty. Further details regarding these steps are provided in Appendix B.

3.3.2 Human Verification

To ensure the quality of our constructed dataset, we conducted human verification. Specifically, we randomly sampled 300 instances and asked three independent annotators to assess the validity of the generated erroneous reasoning steps. The annotation process yielded an agreement rate exceeding 95%, confirming the high quality of our dataset. Further details regarding the annotators and the annotation process are provided in Appendix C.

4 Experiments

To provide a comprehensive evaluation of various models on MPBENCH, we selected a diverse set, encompassing both open-source and proprietary LLMs. This selection includes advanced models such as GPT-4 and multi-step reasoning-enhanced LLMs like the Gemini-2-Thinking (DeepMind, 2024).

Given the complexity of the tasks, few-shot demonstration setups were employed to facilitate model adaptation to the required output format through In-Context Learning (ICL). We utilized two-shot examples when prompting general-purpose LLMs. The impact of varying few-shot settings is discussed in Section 4.1.4. The prompts for the above phases are detailed in Appendix E.

4.1 Results and Analysis

Table 2 showcases the performance of MLLMs on MPBench. Specifically, we are interested in exploring the following research questions: RQ1: How do model architecture and scale impact performance? RQ2: What correlations exist between different reasoning abilities (step correctness identity, answer aggregation, reasoning process search)? RQ3: How does performance vary across different steps in a multi-step reasoning process? RQ4: What is the effect of in-context learning (ICL) settings on the model's performance? RQ5: How does domain knowledge (science, mathematics, commonsense) influence performance? In the following sections, we will discuss these research questions in turn.

4.1.1 RQ1: Influence of Model & Scale

Model As shown in Fig 2, they fail to generalize across the different tasks, especially on the First Error Identification (FEI) of step correctness and

Model Name	Overall	Step Correctness			Answer Aggregation			Reasoning Process Search		
		FEI	AEI	Avg.	BoN	MV	Avg.	F1	MCC	Avg.
Random	31.8	9.1	50.0	29.6	40.9	40.9	40.9	50.0	0	25.0
InternVL2.5-1B	31.3$_{-0.5}$	22.2	49.1	35.6	47.1	33.3	40.2	45.5	-9.1	18.2
InternVL2.5-8B	48.3$_{+16.5}$	36.7	56.9	46.8	79.5	58.0	68.7	53.1	6.3	29.7
InternVL2.5-26B	42.4$_{+10.6}$	10.3	51.7	31.0	67.5	55.6	61.5	56.5	12.9	34.7
InternVL2.5-38B	49.7$_{+17.9}$	55.3	28.2	38.9	74.6	57.2	66.9	55.4	11.2	33.5
Qwen2.5-VL-3B	37.6$_{+5.8}$	1.6	48.2	24.9	76.8	55.0	65.9	48.1	-3.8	22.1
Qwen2-VL-7B	42.4$_{+10.6}$	6.8	51.3	29.1	85.2	50.6	67.9	53.6	7.1	30.3
Qwen2.5-VL-7B	45.7$_{+13.9}$	9.8	54.2	32.0	59.0	52.5	55.8	66.3	32.5	49.4
Qwen2.5-VL-72B	55.8$_{+24.0}$	24.8	65.2	45.0	81.6	**69.8**	75.7	64.4	28.8	46.6
QVQ	41.8$_{+10.0}$	15.5	55.6	35.5	63.6	51.7	57.7	54.8	9.6	32.2
VisualPRM	-	-	-	-	83.0	-	-	-	-	-
Claude-3.7	**71.2**$_{+39.4}$	50.8	71.8	61.3	**85.8**	66.4	**76.1**	77.0	58.6	67.8
Gemini-2.0-flash-exp	65.4$_{+33.6}$	53.8	**74.9**	**64.4**	82.2	69.6	75.9	70.5	41.1	55.8
Gemini-2.0-thinking-exp	64.6$_{+32.8}$	**54.0**	**74.9**	**64.4**	81.7	67.7	74.7	69.8	39.6	54.7
GPT-4o	**71.2**$_{+39.4}$	51.2	74.4	62.8	85.3	62.2	73.7	**84.7**	**69.5**	**77.1**

Table 2: Performances comparison of models on MPBench. The best performance for each category and task is in **bold**, while the second-best performance is underlined. Random denotes the performance of random reward generation. In the Overall category, + (or −) indicates the performance gain compared to the random baseline.

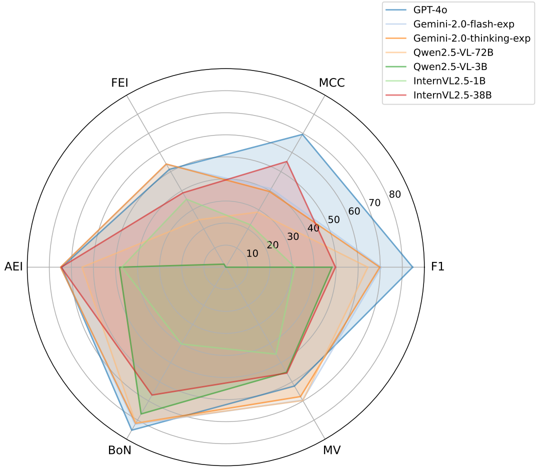


Figure 2: Performance breakdown on MPBench.

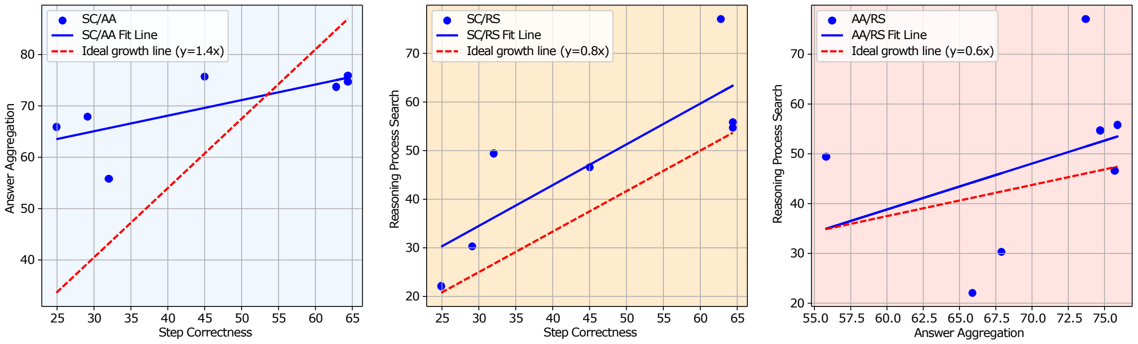
Majority Voting (MV) of answer aggregation, indicating potential limitations in their ability to effectively process and integrate multimodal information for error detection and answer selection. On the other hand, their performance on All Error Identification (AEI) and Best-of-N (BoN) is significantly higher, suggesting that directly leveraging LLMs for reward process reasoning during both training and inference might be more effectively achieved through AEI and BoN. This discrepancy in performance across different evaluation paradigms raises interesting questions about the optimal design and utilization of PRMs for multimodal reasoning tasks.

Among the overall metrics, the state-of-the-art model GPT-4o performs the best. While consistently maintaining comparability, its performance in Reasoning Process Search (F1 and MCC) is significantly higher than existing open-source and proprietary models, indicating that stronger model capabilities are required for reward models to search paths during the inference stage.

Scale Model performance on MPBench generally scales with size, most notably for Step Correctness and Reasoning Process Search (Table 2). Weaker models (e.g., Qwen2.5-VL-3B) even perform below random chance on these assessments. This suggests that larger model capacity is crucial for complex reasoning, enabling better learning of correct/incorrect steps and navigation of the solution space. The disproportionate impact of scale on Step Correctness and Search suggests these tasks are cognitively demanding, requiring deeper reasoning process understanding and step-level evaluation.

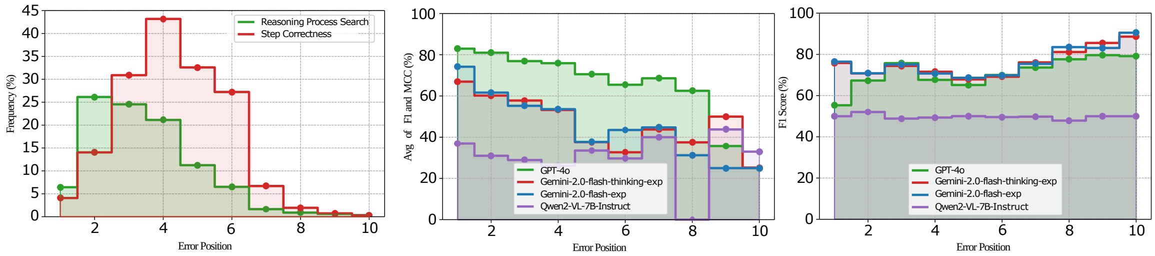
4.1.2 RQ2: Correlations of different Abilities

The capabilities of step correctness, answer aggregation, and reasoning process search show a posi-



(a) Step Correctness vs Answer Aggregation (b) Step Correctness vs Reasoning Process Search (c) Answer Aggregation vs Reasoning Process Search

Figure 3: Interrelationship between a model's capabilities in step correctness identify, answer aggregation, and reasoning process search. Each point on the graph represents a model, with coordinates indicating its performance in step correctness identify(SC), answer aggregation (AA), and reasoning process search (RS). The graph features fitted lines for the scatter plots, denoted by blue lines for SC/AA, SC/RS, and AA/RS, while a red dashed line represents the ideal growth line. The slope of this ideal growth line is the ratio of the random values of each metric.



(a) Distribution of error positions (b) Reasoning Process Search (c) Step Correctness

Figure 4: Impact of Error Position on Model Performance. (a) Distribution of error positions within the dataset. (b) Model performance on reasoning process search, measured by average F1 score and MCC, across different error positions. (c) Model performance on Step Correctness, measured by F1 score, across different error positions. Note: Step 1 and steps beyond 10 are truncated for improved visualization.

tive correlation. Fig 3 illustrates the interrelationship among these capabilities. A positive linear relationship is observed between step correctness and reasoning process search, with the improvement rates in both generation and critique being nearly identical, despite step correctness focusing primarily on training parsing. However, the linear correlation between step correctness and answer aggregation is less pronounced. Although increases in step correctness lead to improvements in answer aggregation, the growth rate of answer aggregation is slower, suggesting that its enhancement requires targeted exploration beyond simply improving step correctness. As seen in Fig 3 (c), there is a notable increase in reasoning process search as answer aggregation improves, indicating that reasoning process search benefits from enhanced answer aggregation capabilities.

4.1.3 RQ3: Performance at different step positions

Fig 4 (a) displays the distribution of error positions within the dataset. As illustrated, the highest concentration of errors occurs at step 4, with a gradual decline in frequency towards both earlier and later steps. While errors are present across all steps shown (2 through 10), the distribution is clearly skewed towards the mid-range.

The performance of different models on reasoning process search and step correctness at various error positions is shown in Figures 4(b) and (c), respectively. For reasoning process search (b), we observe a general downward trend in performance as the error position increases. This suggests that models may struggle to effectively navigate the search space as the reasoning process lengthens, potentially due to the accumulation of errors or the exponential expansion of possible paths. In contrast, the performance on Step Correctness (c) appears less sensitive to the error position and, in some cases, even shows improvement in later steps. This indicate that models become more confident in their step evaluations with more context or that training data might be biased toward later steps.

Comparing the two evaluation paradigms, we

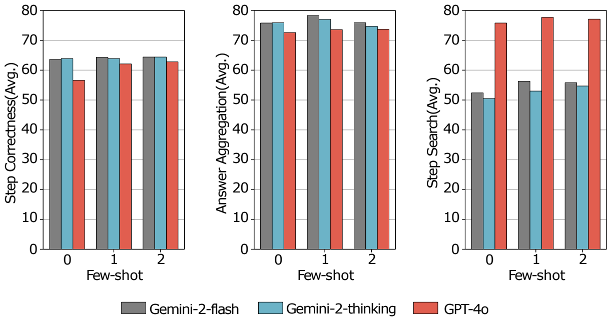


Figure 5: The impact of ICL few-shot numbers on model performance.

Evaluation	Model	Science	Commonsense	Mathematics
Step Correctness R:29.6	Qwen2-VL-7B-Instruct	$29.0_{-0.6}$	$31.3_{+1.7}$	$29.2_{-0.4}$
	Gemini-2-thinking	$67.3_{+37.7}$	$64.7_{+35.1}$	$38.2_{+8.6}$
	GPT-4o	$63.1_{+33.5}$	$66.2_{+36.6}$	$46.3_{+16.7}$
Answer Aggregation R:40.9	Qwen2-VL-7B-Instruct	$66.3_{+26.9}$	$71.2_{+30.3}$	$68.5_{+27.6}$
	Gemini-2-thinking	$77.8_{+36.9}$	$68.6_{+27.7}$	$72.7_{+31.8}$
	GPT-4o	$75.7_{+34.8}$	$72.9_{+32.0}$	$68.6_{+27.7}$
Process Search R:25.0	Qwen2-VL-7B-Instruct	$29.3_{+4.3}$	$32.7_{+7.7}$	$24.3_{-0.7}$
	Gemini-2-thinking	$53.1_{+28.1}$	$59.4_{+34.4}$	$42.9_{+17.9}$
	GPT-4o	$77.8_{+52.8}$	$77.5_{+52.5}$	$70.0_{+45.0}$

Table 3: Performance of Models on MPBench Across Science, Commonsense, and Mathematics Domains. The table presents average scores for each evaluation paradigm (Step Correctness, Answer Aggregation, and Reasoning Process Search), compared to the performance of a random baseline (R).

can see a clear divergence in their performance patterns across different error positions. While reasoning process search emphasizes the ability to identify the correct path early in the reasoning process, step correctness focuses on evaluating the correctness of individual steps, regardless of their position. This suggests that these two paradigms capture different aspects of reasoning ability and may be relevant at different stages of the reasoning.

4.1.4 RQ4: Impacts of ICL settings

Fig 5 presents the evaluation of GPT-4o, Gemini-2-flash, and Gemini-2-thinking across varying ICL few-shot conditions (0, 1, and 2 shots). For the Gemini models, a marginal performance improvement is observed with the introduction of few-shot learning. Specifically, both Gemini models show slight gains in step correctness and answer aggregation when moving from zero-shot to few-shot prompting. GPT-4o's performance, particularly in Step Correctness, demonstrates a more pronounced improvement between the zero-shot and one-shot settings. However, further increases in the few-shot count (from 1 to 2) do not consistently yield additional performance gains and, in the case of Answer Aggregation for GPT-4o, even correlate with a marginal decline. Overall, these findings suggest that the impact of ICL few-shot learning on MPBench performance is subtle. While some benefit is apparent, particularly in the transition from zero-shot to one-shot, the effect is not substantial, indicating that the models may be relatively robust to the number of demonstrations or that the provided demonstrations do not offer significant supplementary information for these tasks.

4.1.5 RQ5: Impacts of Domain

Table 3 presents the results across different domains. The questions are categorized into three primary domains: science knowledge, mathematics, and commonsense, consistent with the categories already identified in M3CoT. As shown in the table, performance generally declines in the mathematics domain, which is arguably more challenging. Notably, GPT-4o exhibits a significant performance advantage in this category. This suggests that GPT-4o's stronger capabilities are more effectively leveraged when dealing with complex, mathematically oriented problems within the process reward framework. In contrast, Qwen2-VL-7B-Instruct's performance in mathematics falls below the random baseline, highlighting the difficulties faced by less capable models in this domain. This underscores the importance of robust model capacity for effectively utilizing process rewards, particularly when dealing with challenging problem domains. The performance differences across domains suggest that future research could benefit from developing domain-specific PRMs or training strategies that better equip models to handle diverse reasoning demands. Furthermore, the substantial gap between GPT-4o and other models in the mathematics domain indicates that this area remains a significant challenge and a promising direction for future advancements in process reward modeling.

5 Conclusions

This work presents MPBench, a novel multimodal benchmark designed for evaluating error identification in reasoning processes. Through a detailed analysis of MLLM performance on MPBench, we have shed light on the strengths and weaknesses of these models when used as criteria for enhancing reasoning. Specifically, we examined three key aspects: step correctness identification, answer aggregation, and reasoning process search. Our investigation revealed a linear correlation coupled with subtle inconsistencies between them. Additionally, our analysis across diverse domains (science, commonsense, and mathematics) demonstrated that mathematical reasoning presents a significant challenge for current MLLMs, with performance lagging behind that observed in other domains. This suggests that future research should prioritize the development of PRMs and training methodologies that are more robust to the complexities of mathematical reasoning.

Limitations

Despite our best efforts throughout the entire benchmark construction process, MPBench may still contain inaccurate labels of error locations, particularly for the more challenging complex math problems.

Acknowledgement

This paper is supported by the National Key R&D Program of China No.2022ZD0160102, the National Natural Science Foundation of China (62476069, 62441202 and 62176165) and the Stable Support Projects for Shenzhen Higher Education Institutions (20220718110918001). This work was done during his internship at Shanghai Artificial Intelligence Laboratory.

References

Jiaxin Ai, Pengfei Zhou, Zhaopan Xu, Ming Li, Fanrui Zhang, Zizhen Li, Jianwen Sun, Yukang Feng, Baojin Huang, Zhongyuan Wang, et al. 2025. Projudge: A multi-modal multi-discipline benchmark and instruction-tuning dataset for mllm-based process judges. *arXiv preprint arXiv:2503.06553*.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024a. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint arXiv:2405.16473*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. 2024. Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. *arXiv preprint arXiv:2412.12932*.

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, 2021. *URL https://arxiv.org/abs/2110.14168*.

DeepMind. 2024. Gemini 2.0 flash experimental. `https://deepmind.google/technologies/gemini/flash/`. Accessed: 2024-12-25.

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*.

Hyeonbin Hwang, Doyoung Kim, Seungone Kim, Seonghyeon Ye, and Minjoon Seo. 2024. Self-explore to avoid the pit: Improving the reasoning capabilities of language models with fine-grained rewards. *arXiv preprint arXiv:2404.10346*.

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*.

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. 2024. Criticbench: Benchmarking llms for critique-correct reasoning. *arXiv preprint arXiv:2402.14809*.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

OpenAI. 2024. Learning to reason with llms. `https://openai.com/index/learning-to-reason-with-llms/`.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. 2025. Prmbench: A fine-grained and challenging benchmark for process-level reward models. *arXiv preprint arXiv:2501.03124*.

Qwen Team. 2024. Qvq: To see the world with wisdom.

Qwen Team. 2025. Qwen2.5-vl.

Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, Hisham Cholakkal, Ivan Laptev, Mubarak Shah, Fahad Shahbaz Khan, and Salman Khan. 2025. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *Preprint*, arXiv:2501.06186.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2023. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, et al. 2025. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. 2024. Evaluating mathematical reasoning beyond accuracy. *arXiv preprint arXiv:2404.05692*.

Kun Xiang, Zhili Liu, Zihao Jiang, Yunshuang Nie, Runhui Huang, Haoxiang Fan, Hanhui Li, Weiran Huang, Yihan Zeng, Jianhua Han, et al. 2024. Atomthink: A slow thinking framework for multimodal mathematical reasoning. *arXiv preprint arXiv:2411.11930*.

Zhaopan Xu, Pengfei Zhou, Weidong Tang, Jiaxin Ai, Wangbo Zhao, Xiaojiang Peng, Kai Wang, Yang You, Wenqi Shao, Hongxun Yao, et al. 2025. Pebench: A fictitious dataset to benchmark machine unlearning for multimodal large language models. *arXiv preprint arXiv:2503.12545*.

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.

Zhongshen Zeng, Pengguang Chen, Shu Liu, Haiyun Jiang, and Jiaya Jia. 2023. Mr-gsm8k: A meta-reasoning benchmark for large language model evaluation. *arXiv preprint arXiv:2312.17080*.

Zhongshen Zeng, Yinhong Liu, Yingjia Wan, Jingyao Li, Pengguang Chen, Jianbo Dai, Yuxuan Yao, Rongwu Xu, Zehan Qi, Wanru Zhao, Linling Shen, Jianqiao Lu, Haochen Tan, Yukang Chen, Hao Zhang, Zhan Shi, Bailin Wang, Zhijiang Guo, and Jiaya Jia. 2024. Mr-ben: A meta-reasoning benchmark for evaluating system-2 thinking in llms. *CoRR*, abs/2406.13975.

Wangbo Zhao, Yizeng Han, Jiasheng Tang, Kai Wang, Yibing Song, Gao Huang, Fan Wang, and Yang You. 2024a. Dynamic diffusion transformer. *arXiv preprint arXiv:2410.03456*.

Wangbo Zhao, Jiasheng Tang, Yizeng Han, Yibing Song, Kai Wang, Gao Huang, Fan Wang, and Yang You. 2024b. Dynamic tuning towards parameter and inference efficiency for vit adaptation. *arXiv preprint arXiv:2403.11808*.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2024. Processbench: Identifying process errors in mathematical reasoning. *arXiv preprint arXiv:2412.06559*.

Zihao Zhou, Shudong Liu, Maizhen Ning, Wei Liu, Jindong Wang, Derek F Wong, Xiaowei Huang, Qiufeng Wang, and Kaizhu Huang. 2024. Is your model really a good math reasoner? evaluating mathematical reasoning with checklist. *arXiv preprint arXiv:2407.08733*.

Appendix

A Statistics of our benchmark

Evaluation	Science	Commonsense	Mathematics	All
Step Correctness	2248	1133	322	3703
Answer Aggregation	1535	496	437	2468
Reasoning Process Search	2232	973	369	3574

Table 4: Statistics of MPBench.

B Filtering

Step 1: Rule-Based Filtering for Format Adherence To ensure dataset consistency and usability, we implemented rule-based filtering, defining explicit rules for data instance format (question structure, solution format, metadata). Automated scripts flagged and removed instances violating these rules, ensuring uniformity and preparing the data for further quality control.

Step 2: GPT-4 Review with In-Context Examples. Following the rule-based filtering, we employed a more nuanced approach using GPT-4, to identify and remove unreasonable or nonsensical instances. To effectively guide GPT-4 in this task, we manually curated a set of in-context (IC) examples. These examples consisted of both valid and invalid data instances, carefully chosen to illustrate the types of issues we were looking to identify, such as logical inconsistencies, factual errors, or nonsensical reasoning steps. We then presented GPT-4 with the remaining data instances and asked it to classify each instance as either "reasonable" or "unreasonable" based on the patterns observed in the IC examples.

Step 3: Gemini-Based Difficulty Filtering To ensure dataset difficulty, we used Gemini-1.5-pro to score solution steps, calculating the absolute score difference between incorrect and correct steps. Instances with a difference > 1 (deemed potentially too easy) were manually reviewed and removed if necessary. This multi-stage filtering process, combining automated rules, LLM review, and difficulty assessment, yielded a high-quality dataset suitable for evaluating and training process reward models.

C Human Verification

To ensure dataset quality, we conducted human verification of 300 randomly sampled instances (100 per evaluation paradigm). Three independent annotators, including one co-author and two undergraduate volunteers, assessed the validity of generated erroneous reasoning steps. Annotators reviewed original data, generated erroneous steps, and reasoning context, judging plausibility and alignment with expected error types. A >95% agreement rate across paradigms confirms the dataset's high quality and suitability for evaluating and training process reward models. Minor discrepancies likely reflect the subjective nature of error evaluation in complex reasoning.

D Correlating with Policy Models

Math-Shepherd (Wang et al., 2023) demonstrates that Process Reward Models (PRMs) can significantly enhance Large Language Model (LLM) Reinforcement Learning for reasoning tasks by accurately identifying error locations within the reasoning process. However, the existing challenges include the time-consuming nature of process-supervised training and the lack of an open-source PRM framework specifically designed for Multimodal Large Language Models (MLLMs). To address these gaps and validate the utility of PRMs, our validation experiments specifically focus on two critical aspects: 1) Answer Aggregation and 2) Reasoning Process Search.

1) Answer Aggregation. For Answer Aggregation, we adopted the methodology presented in VisualPRM (Wang et al., 2025). In this approach, policy models are tasked with generating multiple candidate solutions (specifically, 8 distinct Chain-of-Thought (CoT) reasoning processes with a temperature of 0.7 from Qwen2.5-VL-7B). Subsequently, critic models then utilize either the Best-of-N (BoN) or Majority Voting (MV) strategy to aggregate these candidates and determine the final answer.

No.	Model	MMMU	MathVista
#1	Qwen2.5-VL-7B	55.0	67.8
#2	+ Qwen2.5-VL-3B MV	56.8 ↑1.8	68.5 ↑0.7
#3	+ Qwen2.5-VL-3B BoN	58.2 ↑3.2	69.8 ↑2.0
#4	+ Qwen2.5-VL-72B MV	57.3 ↑2.3	69.0 ↑1.2
#5	+ Qwen2.5-VL-72B BoN	57.9 ↑2.9	70.1 ↑2.3

Table 5: Answer Aggregation. Datasets: MMMU (Yue et al., 2024) and MathVista (Wang et al., 2024). Policy Model: Qwen2.5-VL-7B, generating 8 distinct Chain-of-Thought (CoT) reasoning processes (temperature 0.7). Answer Aggregation Strategies: Best-of-N (BoN) and Majority Voting (MV). For example, "+Qwen2.5-VL-3B BoN" signifies Qwen2.5-VL-3B acting as the critic model employing the BoN strategy.

Analysis & Conclusion: Our findings demonstrate that the Best-of-N (BoN) strategy consistently outperforms Majority Voting (MV), which aligns with the observations from MPBench's own findings regarding answer aggregation. Furthermore, experiments with larger-scale MLLMs (models 4 and 5) did not yield significant performance gains. This result corroborates our broader observation that simply using a larger MLLM as a critic does not inherently lead to improved performance in Answer Aggregation tasks.

2) Reasoning Process Search For Reasoning Process Search, we employed AtomThink's methodology (Xiang et al., 2024). In this setup, a policy model is dynamically prompted to generate single-step candidates. Subsequently, critic models play a crucial role in selecting the most appropriate steps for iterative inference, guiding the overall reasoning process.

No.	Model	MMMU	MathVista
#1	Qwen2.5-VL-7B	55.0	67.8
#2	+ Qwen2.5-VL-3B	56.2 ↑1.2	67.9 ↑0.1
#3	+ GPT-4o	58.3 ↑3.3	69.4 ↑1.6

Table 6: Reasoning Process Search. Datasets: MMMU (Yue et al., 2024) and MathVista (Wang et al., 2024). Policy Model: Qwen2.5-VL-7B, generating 5 candidate pre-steps. Critic Models: Select steps for subsequent iterative inference.

Analysis & Conclusion: The results reveal a notable performance gap in Reasoning Process Search between open-source models and closed-source models like GPT-4o. This observation is consistent with the insights gleaned from MPBench's comprehensive evaluations. The findings underscore that Reasoning Process Search is a demanding task that necessitates the use of high-performance critic models to effectively guide the iterative inference process and achieve substantial improvements in reasoning accuracy.

E Prompts

As introduced in Section 3, MPBench employs three evaluation paradigms to assess the capabilities of PRMs in the reasoning process: (1) Step Correctness, (2) Answer Aggregation, and (3) Reasoning Process Search. To evaluate these capabilities, we carefully designed prompts to query MLLMs (e.g., GPT-4o) and assess their performance as PRMs. Below, we provide an example prompt for each paradigm. Due to space limitations, we display only one example for each paradigm.

E.1 Prompts for Step Correctness

As discussed in Section 3.2.1, the Step Correctness paradigm evaluates PRMs' ability to assess the correctness of each intermediate reasoning step. To test this capability, we designed the few-shot prompt as shown in Table 7.

E.2 Prompts for Answer Aggregation

The Answer Aggregation paradigm, as introduced in Section 3.2.2, examines PRMs' ability to aggregate scores from multiple solutions and select the best candidate response. To evaluate this capability, we used the prompt detailed in Table 8.

E.3 Prompts for Reasoning Process Search

Finally, the Reasoning Process Search paradigm, described in Section 3.2.3, evaluates PRMs' ability to guide the search for optimal reasoning steps during inference. To assess this capability, we employed the prompt as in Table 9.

1. System Prompt

You are a reasoning evaluator. Your task is to analyze problem-solving steps and provide structured assessments in JSON format.

For each solution step, you need to evaluate:
Score (-1 to +1):
* +1: Completely correct reasoning
* 0: Partially correct with some mistakes
* -1: Completely incorrect
* Use any value in between to indicate varying degrees of correctness

Requirements:
- Evaluate each step independently
- Provide scores as floating-point numbers
- Return results in strict JSON format: "Score": [scores]
- Ensure arrays have the same length with the number of solution steps
- Consider logical accuracy, mathematical coherence, and solution efficiency

Example output format:
{"Score": [0.8, -0.5, 1.0]}

You will be presented with a problem, its step-by-step solution and its final answer. Please analyze each step and provide your evaluation in the specified JSON format.

2. Few Shots

User 1

Question:

[Question]
What might be the possible function of the area?
[Choices]
(A) Putting on makeup
(B) Conducting business meetings
(C) Taking a rest
(D) Displaying artwork

Solution:

Step 1. We can infer that the area closed off by the long red rope fence in the museum exhibit is used to display artwork as there is a large painting hanging on the wall to the left of the bed with a golden drape, which suggests that the person who owned this place loved artwork.
Step 2. Additionally, the very large bed with a long golden blanket draped around it on each side in a dimly lit room is a luxury item indicating that the person who owned it was wealthy, further supporting the idea that the area is a museum exhibit.
Step 3. The lights in the area may serve as spotlights to focus on the artwork.
Step 4. The lights is suit for conducting business meetings
Step 5. Therefore, we can conclude that the correct answer is D, displaying artwork

Answer:

D

Assistant 1

{"Score": [1.0, 0.7, 0.8, -1.0, 1.0]}

User 2

Question:

[Question]
Which property do these three objects have in common?

[Choices]
(A) scratchy
(B) flexible
(C) fragile

Solution:

Step 1: Examine each object.

Step 2: Determine if each object possesses the specified property.

Step 3: A flexible object is capable of being folded or bent without breaking easily. None of the objects meet this criteria.

Step 4: A scratchy object feels rough and causes itchiness when it comes into contact with the skin. None of the objects fit this description.

Step 5: A fragile object will shatter into multiple pieces if dropped. However, none of the objects are fragile.

Step 6: All three objects are actually flexible.

Step 7: Therefore, while previously concluded as option C, the steps now imply a contradiction indicating they are flexible.

Answer:

C

Assistant 2

{"Score": [1, 0.6, 0.5, 0.8, -1, -1, 1]}

Table 7: Few-shot Prompt for evaluating PRM's ability to assess the correctness of each intermediate reasoning step.

---

1. System Prompt
You are a reasoning evaluator. Your task is to analyze problem-solving steps and provide structured assessments in JSON format.

For each solution step, you need to evaluate:
Score (-1 to +1):
* +1: Completely correct reasoning
* 0: Partially correct with some mistakes
* -1: Completely incorrect
* Use any value in between to indicate varying degrees of correctness

Requirements:
- Evaluate each step independently
- Provide scores as floating-point numbers
- Return results in strict JSON format: "Score": [scores]
- Ensure arrays have the same length with the number of solution steps
- Consider logical accuracy, mathematical coherence, and solution efficiency

Example output format:
{"Score": [0.8, -0.5, 1.0]}

You will be presented with a problem and its step-by-step solution. Please analyze each step and provide your evaluation in the specified JSON format.
2. Few Shots
User 1
Question:
[Question]
What might be the possible function of the areay?
[Choices]
(A) Putting on makeup
(B) Conducting business meetings
(C) Taking a rest
(D) Displaying artwork
Solution:
Step 1. We can infer that the area closed off by the long red rope fence in the museum exhibit is used to display artwork as there is a large painting hanging on the wall to the left of the bed with a golden drape, which suggests that the person who owned this place loved artwork.
Step 2. Additionally, the very large bed with a long golden blanket draped around it on each side in a dimly lit room is a luxury item indicating that the person who owned it was wealthy, further supporting the idea that the area is a museum exhibit.
Step 3. The lights in the area may serve as spotlights to focus on the artwork.
Step 4. The lights is suit for conducting business meetings
Step 5. Therefore, we can conclude that the correct answer is D, displaying artwork
Assistant 1
{"Score": [1.0, 0.7, 0.8, -1.0, 1.0]}
User 2
Question:
[Question]
If you had to select one option as the correct answer for the Precalculus problem shown in the picture, which one would you choose and why?
[Choices]

(A) None of the choices given in the text are correct.
(B) the answer is option D in the image
(C) option B in the image should be selected
(D) the answer is option A in the image

Solution:

Let A $= (\alpha, 0, 0)$, B $= (0, \beta, 0)$, and C $= (0, 0, \gamma)$.

Then the equation of plane ABC is given by $\frac{x}{\alpha} + \frac{y}{\beta} + \frac{z}{\gamma} = 1$.

Assuming the relationship between the coordinates is incorrect, we might hypothesize a different approach where $p = \frac{\alpha}{3}$,$q = \frac{\beta}{3}$,and $r = \frac{\gamma}{3}$ gives $\frac{1}{p^2} + \frac{1}{q^2} + \frac{1}{r^2} = \frac{3}{\alpha^2} + \frac{3}{\beta^2} + \frac{3}{\gamma^2}$.

By equating algebraic blunders, assume $\frac{3}{\alpha^2} + \frac{3}{\beta^2} + \frac{3}{\gamma^2} = 3$, without correct reasoning.

Then, the incorrect conclusion is $\boxed{1}$, misreading the relations and not equating correctly.

Determine the incorrect answer based on the flawed assumption in the provided image Choices: (A) 9, (B) 5, (C) 36, (D) 1

Based on incorrectly assumed calculations, the answer would be option D in the image.

Consequently, due to faulty reasoning, the incorrect solution is choice (B) in the given text.

Assistant 2

{"Score": [1.0, 0.6, 0.3, -1.0, -1.0, 0.0, -1.0]}

Table 8: Few-shot Prompt for evaluating PRM's ability to aggregate scores from multiple solutions and select the best candidate response.

1. System Prompt
Your task is to evaluate the next step of reasoning or calculation based on THE GIVEN QUESTION and HISTORICAL REASONING STEPS.

You will be provided with:
1. A QUESTION.
2. HISTORICAL REASONING STEPS.
3. Candidates of next step.

For each solution step, you need to evaluate:
Score (-1 to +1):
* +1: Completely correct reasoning
* 0: Partially correct with some mistakes
* -1: Completely incorrect
* Use any value in between to indicate varying degrees of correctness

Requirements:
- Evaluate each candidate independently
- Provide scores as floating-point numbers
- Return results in strict JSON format: "Score": [scores]
- Ensure arrays have the same length with the number of candidates
- Consider logical accuracy, mathematical coherence, and solution efficiency

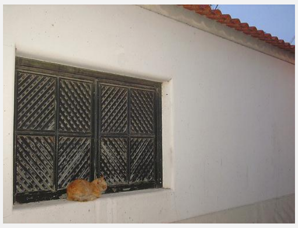
Example output format:
{"Score": [0.8, -0.5]}
2. Few Shots
User 1
Question:
[Question]
What kind of snowboarders is the mountain in the picture suitable for?
[Choices]
(A) Beginner only
(B) Not sure
(C) Advanced only
(D) All levels
Historical reasoning steps:
Step 1. The snowboarder in the picture is seen touching the ground with his arm.
Step 2. It indicates that they are struggling to maintain their balance.
Candidates of next step:
So it suggests that the mountain is challenging for the man in the image.
However, the snowboarder might also be intentionally leaning to perform a trick, which some beginners may attempt, but it could also imply a deceptive appearance of difficulty.
Assistant 1
{"Score": [0.8, -0.7]}
User 2
Question:
[Question]
Which property do these four objects have in common?
[Choices]
(A) salty
(B) stretchy

(C) transparent
Historical reasoning steps:
Step 1: Examine each object individually.
Step 2: Determine if each object possesses the specified property.
Step 3: A stretchy object elongates when force is applied. However, the potato chips and the pretzel are not stretchy.
Step 4: Potato chips are known for their salty taste. It is important to note that all four objects are salty.
Candidates of next step:
Step 5: While a transparent object allows clear visibility through it, note that ocean water can appear transparent under certain conditions. Consequently, only the ocean water might be considered transparent, while the potato chips, the pretzel, and the fries are not.
Step 5: A transparent object allows clear visibility through it. However, the potato chips, the pretzel, and the fries are not transparent.
Assistant 2
{"Score": [-0.5, 0.6]}

Table 9: Few-shot Prompt for evaluating PRM's ability to guide the search for optimal reasoning steps during inference.

Question:

What inference can be made from the combination of the following clues: an orange cat sitting in a windowsill in the shade and a white stucco wall on a building?



A. "The cat is trying to avoid direct sunlight",
B. "The cat is a stray looking for a home",
C. "The cat enjoys lounging in the sun",
D. "The building is located in a cold environment"

Solution process:

1. "We can infer from the image of the orange cat sitting in the windowsill in the shade that the cat is trying to cool down.",
2. "Additionally, from the image of the white stucco wall on the building, we can infer that the building is in a hot environment.",
3. "Therefore it would make sense that the cat is trying to avoid direct sunlight and is sitting in the shade.",
4. "However, if the cat is sitting in the shade, it is likely basking in the sun.",
5. "Option D is incorrect as we are told that the building has a white stucco wall, which typically absorbs heat and indicates the building is in a hot environment.",
6. "Therefore, option A is the correct answer."

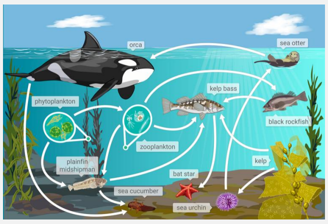
Label: [1, 1, 1, 0, 1, 1]

Figure 6: **Erroneous Steps.**

Question:

In this food web, which organism contains matter that eventually moves to the bat star?

A. "sea cucumber",
B. "sea otter",
C. "phytoplankton"



Solution process 1: √

1. "From the phytoplankton, there are four potential paths: a) Phytoplankton –> Zooplankton –> Kelp Bass –> Bat Star, b) Phytoplankton –> Zooplankton –> Plainfin Midshipman –> Kelp Bass –> Bat Star, c) Phytoplankton –> Zooplankton –> Black Rockfish –> Kelp Bass –> Bat Star, d) Phytoplankton –> Plainfin Midshipman –> Kelp Bass –> Bat Star.",
2. "From the zooplankton, there are three options: a) Zooplankton –> Kelp Bass –> Bat Star, b) Zooplankton –> Plainfin Midshipman –> Kelp Bass –> Bat Star, c) Zooplankton –> Black Rockfish –> Kelp Bass –> Bat Star.",
3. "Moving on to the sea otter, there is only one arrow pointing from the sea otter, which leads to the orca.",
4. "Following the arrow from the orca, it leads to the sea cucumber.",
5. "However, no arrows point from the sea cucumber to any other organisms. Therefore, matter does not move from the sea cucumber to the bat star."

Solution process 2: ✗

1. "From the phytoplankton, we first check the direct paths. Although they seem to lead to the bat star, we find that there is actually a different indirect link.",
2. "Ignoring the paths to kelp bass, I propose that zooplankton might directly lead to the orca, suggesting they could be indirectly related to the bat star through another food chain.",
3. "Moving to the sea otter, I state that it also contributes matter to the sea cucumber rather than going to the bat star.",
4. "Therefore, I conclude that the sea cucumber is directly linked to the bat star through the orca, making it a viable answer."

Solution process 3: ✗

1. "The question asks which organism's matter *eventually* moves to the bat star. This implies a longer chain is preferable.",
2. "The sea otter eats kelp bass, and kelp bass can be eaten by a bat star. Therefore, matter from the sea otter could end up in a bat star (sea otter –> kelp bass –> bat star implied).",
3. "While phytoplankton is also a correct answer, the sea otter path involves fewer steps. Since the question doesn't specify the shortest path, the sea otter is a potentially confusing but logically sound answer."

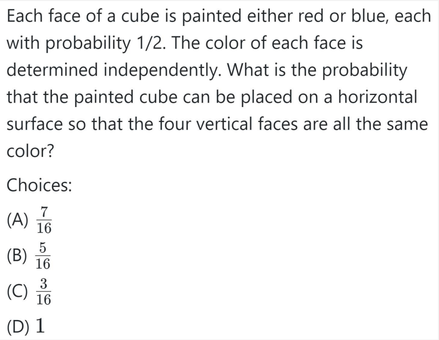
Label: [1, 0, 0]

Figure 7: **Multi-solution.**

Question:

What is the probability that the painted cube can be placed on a horizontal surface so that the four vertical faces are all the same color?

A. "Option B in the image",
B. "Option C in the image",
C. "Option A in the image",
D. "Since all temperatures are above room temperature (25\u00b0C), there will be no dew forming on any object's surface."



Each face of a cube is painted either red or blue, each with probability 1/2. The color of each face is determined independently. What is the probability that the painted cube can be placed on a horizontal surface so that the four vertical faces are all the same color?

Choices:

(A) $\frac{7}{16}$
(B) $\frac{5}{16}$
(C) $\frac{3}{16}$
(D) 1

Solution process:

1. "If the orientation of the cube is fixed, there are $2^6 = 64$ possible arrangements of colors on the faces.",
2. "There are $2\\binom{6}{6}=2$ arrangements in which all six faces are the same color and $2\\binom{6}{5}=12$ arrangements in which exactly five faces have the same color.",
3. "In each of these cases the cube can be placed so that the four vertical faces have the same color.",
4. "The only other suitable arrangements have four faces of one color, with the other color on a pair of opposing faces.",
5. "Since there are three pairs of opposing faces, there are $2(3)=6$ such arrangements.",
6. "The total number of suitable arrangements is therefore $2+12+6=20$, and the probability is $20/64= \\boxed{\\frac{5}{16}}$.",
7. "Select the option that matches the picture:",
8. "Hence, option B in the image is the answer.",
9. "Overall, choice (B) in the given text is the right answer."

Candidates steps:

6 "The total number of suitable arrangements is therefore $2+12+6=20$, and the probability is incorrectly stated as $20/64= \\frac{3}{16}$ instead of $ \\frac{5}{16}$. This may lead to confusion in the final conclusion.",

6 "The total number of suitable arrangements is therefore $2+12+6=20$, and the probability is $20/64= \\boxed{\\frac{5}{16}}$."

Label: [0, 1]

Figure 8: **Action Trees.**