MMSciBench: Benchmarking Language Models on Multimodal Scientific Problems

Anonymous ACL submission

Abstract

Recent advances in large language models (LLMs) and vision-language models (LVLMs) have shown promise across many tasks, yet their scientific reasoning capabilities remain untested, particularly in multimodal settings. We present MMSciBench, a benchmark for evaluating mathematical and physical reasoning through text-only and text-image formats, with human-annotated difficulty levels, solutions with detailed explanations, and taxonomic mappings. Evaluation of state-of-the-art models reveals significant limitations, with even the best model achieving only **63.77%** accuracy and particularly struggling with visual reasoning tasks. Our analysis exposes critical gaps in complex reasoning and visual-textual integration, establishing MMSciBench as a rigorous standard for measuring progress in multimodal scientific understanding. The code for MMSciBench is open-sourced at GitHub[1], and the dataset is available at Hugging Face[2].

1 Introduction

Scientific reasoning represents a crucial test of artificial intelligence (AI) systems' ability to understand and apply complex concepts, making it essential for developing truly intelligent models (Evans et al., 2023; Liang et al., 2024; Zhang et al., 2023; Truhn et al., 2023; Ma et al., 2024; Sprueill et al., 2023).Recent advancements in LLMs like GPTs (Brown et al., 2020; Achiam et al., 2023) and Llama (Dubey et al., 2024) have significantly transformed the field of natural language processing (NLP). Despite these advances, scientific reasoning remains challenging for these models, facing several key limitations: *(1) Lack of multimodal evaluation*: While LVLMs have emerged as powerful models capable of processing both images and text, existing scientific benchmarks are predominantly text-only, preventing comprehensive assessment of visual-textual reasoning abilities. *(2) Limited domain coverage*: Current scientific datasets either focus too narrowly on individual subjects or too broadly across scientific areas, failing to systematically evaluate understanding of key concepts within specific disciplines. *(3) Insufficient assessment granularity*: Existing benchmarks lack human-annotated difficulty levels and structured taxonomies of scientific concepts, making it challenging to evaluate models' performance across different complexity levels and specific knowledge domains. These limitations create an urgent need for a benchmark that can effectively evaluate both LLMs' and LVLMs' scientific reasoning abilities while addressing these challenges.

To address these challenges, we introduce MMSciBench, a benchmark focused on mathematics and physics that evaluates scientific reasoning capabilities. Our benchmark makes three key contributions: (1) A comprehensive evaluation framework that combines multiple-choice questions (MCQs) and open-ended Q&A problems, designed to test diverse reasoning skills across mathematical and physical domains. (2) A novel multimodal assessment approach incorporating both text-only and text-image formats, enabling direct comparison of models' unimodal versus multimodal reasoning capabilities. (3) A hierarchical taxonomy of scientific concepts with human-annotated difficulty levels, detailed solutions, and explanations for each problem. We conducted extensive experiments using four state-of-the-art LVLMs (including both open-source and proprietary models) on the complete dataset, and two mathematics-specialized LLMs on text-only questions. For consistent evaluation across models, we employed GPT-4o as an automated assessor.

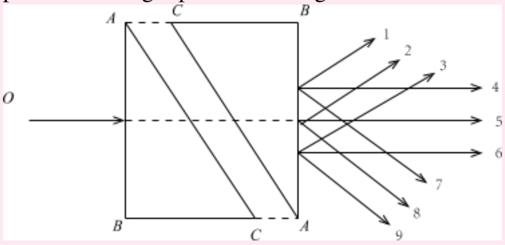
*Corresponding Author.

[1] https://anonymous.4open.science/r/MMSciBench-code-812A/

[2] https://huggingface.co/datasets/anonymous-acl-submission/mmscibench-anonymous

Question & Standard Solution

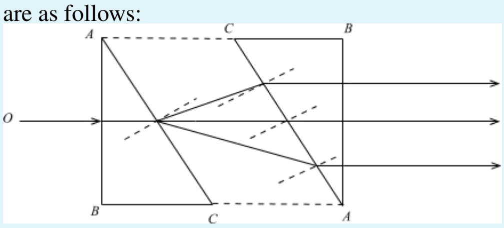
Question
Question (Single Choice): As shown in the figure, two identical right-angled glass prisms ABC are placed with their AC faces parallel to each other, and between them is a uniform unknown transparent medium. A monochromatic thin light beam O is incident perpendicular to the AB face. ( ) is the possible exit light path in the diagram.



Options:
A. Any one of the lines 1, 2, 3 (parallel to each other)
B. Any one of the lines 4, 5, 6 (parallel to each other)
C. Any one of the lines 7, 8, 9 (parallel to each other)
D. Only one of the lines 4 or 6
Standard Solution: B

Explanation

This question primarily tests knowledge of prism-related problems.
Option analysis: According to the problem description, the refractive index of the medium between the two right-angled prisms is unknown. It may be greater than, equal to, or smaller than the refractive index of the glass. The possible light path diagrams are as follows:



Therefore, Option B is correct, and Options A, C, and D are incorrect.
In conclusion, the correct answer to this question is B.

Figure 1: The English translation of an example of a physics MCQ, featuring a single-choice question, the correct answer, and a detailed explanation to aid understanding. The original Chinese version is shown in Fig. 10 in the appendix.

Our evaluation reveals significant limitations in current models' multimodal scientific reasoning capabilities. Gemini 1.5 Pro 002 achieved the highest accuracy (**63.77%**), followed by Claude 3.5 Sonnet (**53.95%**) and GPT-4o (**50.94%**), while Llama-3.2-90B-Vision-Instruct performed substantially lower (**31.19%**). Analysis across task types exposed three critical challenges: (1) Performance degradation on open-ended tasks, with accuracy dropping by an average of **22.32%** compared to multiple-choice questions (2) Systematic failures in complex mathematical and physical reasoning, particularly in domains requiring multi-step problem-solving (3) Limited visual-textual integration, evidenced by a **36.28%** performance gap between text-only and text-image questions Notably, model performance improved when utilizing explicit chain-of-thought prompting and English-language reasoning, even for Chinese-language questions, suggesting potential pathways for enhancing scientific reasoning capabilities.

2 MMSciBench

2.1 Data Collection and Preprocessing

The benchmark data was originally curated by K-12 teachers who annotate questions, detailed step-by-step solutions, final answers, difficulty level, knowledge points, as well as a bunch of other metadata. The dataset[3] includes precise text descriptions, high-resolution images, and high-quality solutions, all compiled and shared as part of a collaborative research effort aimed at advancing AI benchmarking standards. Each question in the dataset is assigned a human-annotated hardness score ranging from 0 to 1, where 1 represents the most challenging questions, and zero denotes the easiest.

To ensure benchmark quality and rigor, we implemented a systematic data curation process. We filtered out questions with incomplete information or duplicate content, focusing on problems with well-defined, quantifiable answers. Following our emphasis on challenging scientific reasoning, we selected questions with human-annotated difficulty scores ≥ 0.7 on a standardized scale. To maintain consistent evaluation conditions, we limited visual content to a maximum of one image per question. To enable systematic knowledge categorization, we employed GPT-4o to annotate each question according to a three-level subject-specific taxonomy, detailed in Section 2.2. The classification results were thoroughly validated by experienced K-12 curriculum specialists to ensure accuracy and alignment with educational standards. This taxonomic analysis confirmed that our filtered dataset maintains comprehensive coverage of key scientific concepts while focusing on challenging problems. Following preprocessing and validation, the final benchmark contains **4,482** question-solution pairs

[3]The dataset is released under the apache-2.0 license.

that enable rigorous evaluation of models' scientific reasoning capabilities across diverse domains.

2.2 Dataset Description

Data Characteristics The MMSciBench dataset offers several distinct advantages over previous scientific datasets:

1. **Curriculum Coverage:** The benchmark spans essential high school mathematics and physics concepts through carefully curated MCQs and open-ended Q&A questions. We maintain comprehensiveness while keeping the dataset size tractable ($N =$ **4,482**).
2. **Quality Assurance:** Questions undergo multi-stage validation by K-12 educators and domain experts, ensuring pedagogical relevance and technical accuracy. Each question includes detailed solutions and explanations.
3. **Multimodal Design:** The parallel text-only and text-image question formats enable systematic comparison of unimodal and multimodal reasoning capabilities.
4. **Structured Assessment:** Questions are organized through a three-level taxonomy and annotated with standardized difficulty scores, facilitating fine-grained analysis of model performance.

An example of a physics MCQ in English is shown in Fig. 1, with the original Chinese version available in Fig. 10 in the appendix. Additionally, a detailed comparison between MMSciBench and other scientific benchmarks is provided from multiple perspectives in Table 1.

Data Statistics MMSciBench comprises **4,482** questions, distributed across modalities and question types, as shown in Table 2. The distribution of core knowledge areas for mathematics and physics is illustrated in Figure 2.

Taxonomy The taxonomy used in MMSciBench has three levels: *Domain*, *Module*, and *Chapter*:

- **Domain**: Core subject areas that define fundamental knowledge boundaries. Mathematics domains include "Sets" and "Functions", while physics encompasses "Classical Mechanics", "Electrodynamics", and "Quantum Mechanics". *Domains* group related topics under a common framework.

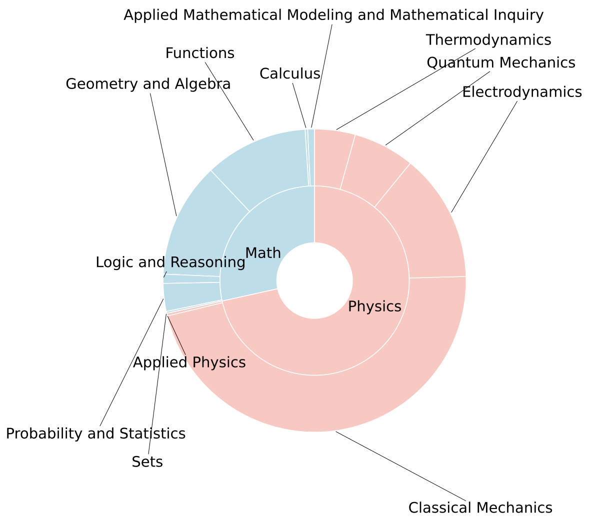


Figure 2: The distribution of data in MMSciBench according to the first-level key knowledge points for each subject.

- **Module**: Subdivisions within *Domains* that focus on key themes or methods. Examples include "Probability and Statistics" in mathematics and "Mechanical Motion and Physical Models" in physics. *Modules* scaffold learning by clustering related topics.
- **Chapter**: The most detailed level, covering specific topics within a *Module*. For instance, mathematics *Chapters* under "Functions" include "Exponential Functions" and "Trigonometric Functions", while physics *Chapters* under "Interactions and Laws of Motion" include "Hooke's Law" and "Equilibrium Conditions of Concurrent Forces". *Chapters* enable fine-grained content analysis and annotation.

3 Experiment Settings

3.1 Evaluated Models

We evaluated our benchmark using four state-of-the-art LVLMs: GPT-4o, Claude 3.5 Sonnet (Anthropic, 2024), Gemini 1.5 Pro 002 (Team et al., 2024), and Llama-3.2-90B-Vision-Instruct.

In addition, as there are models specifically designed for mathematical problem-solving, we extend our evaluation to include two math-focused LLMs: Qwen2.5-Math-72B-Instruct (Yang et al., 2024) and DeepSeekMath-7B-Instruct (Shao et al., 2024). Additionally, we assessed two specialized mathematical LLMs—Qwen2.5-Math-72B-Instruct (Yang et al., 2024) and DeepSeekMath-

Benchmark	Subject	Modality	key knowledge point	Explanation	Language	Difficulty	Size
MSVEC	P, O	T	✗	✓	EN	College	200
SciOL	P, O	T&I	✗	✗	EN	College	18M
TRIGO	M	T	✗	✓	Lean	High School	11K
DMath	M	T	✓	✓	EN&KR	Grade School	10K
GRASP	P	T&V	✓	✗	EN	Basic	2K
SceMQA	M, P, O	T&I	✓	✓	EN	Pre-College	1K
OlympiadBench	M, P	T, T&I	✓	✓	EN, ZH	Olympiad	8K
GAOKAO-Bench	M, P, O	T	✗	✓	ZH	High School	3K
GAOKAO-MM	M, P, O	T, T&I	✗	✓	ZH	High School	650
MMSciBench (Ours)	M, P	T, T&I	✓	✓	ZH	High School	4K

Table 1: Comparison of MMSciBench with existing benchmarks. T denotes text-only data, T&I denotes text-image data pairs, and T&V denotes text-video data pairs. EN, ZH, and KR represent English, simplified Chinese, and Korean, respectively.

Question Type	Math		Physics		Overall	
	MCQs	Q&A	MCQs	Q&A	MCQs	Q&A
Text&Image	260	197	450	260	710	457
Text	500	319	2257	239	2757	558
Total	760	516	2707	499	3467	1015

Table 2: Distribution of questions in MMSciBench by image presence, subject, and question type.

7B-Instruct (Shao et al., 2024)—on the text-only mathematics subset. For reproducibility, all evaluations used a fixed sampling temperature of 0.

3.2 Evaluation Criteria

To evaluate the models, we use accuracy as the metric, a widely adopted standard in existing research, for all question types in MMSciBench. Our evaluation focuses solely on whether the final answer is correct, without considering intermediate solution steps. This criterion is naturally suited for MCQ evaluation, as grading is based on the selected choice(s) in practice. For Q&A questions, this approach ensures a fair and objective comparison by emphasizing the correctness of the final answer rather than incorporating subjective human-defined grading that accounts for intermediate steps.

The evaluation workflow involves first generating answers for MMSciBench questions using each model. GPT-4o is then employed to assess answer correctness by comparing the models' final outputs with the dataset's standard solutions. In existing studies, MCQs often require models to adhere to a specified output format, imposed through prompts, with regular expression rules used to extract the selected choice(s). However, during our experiments, we observed that some models struggled to consistently follow these formatting instructions, complicating this approach. In fact, none of the models achieved a 100% compliance rate with the formatting guidelines. To ensure the evaluation focuses on the models' scientific knowledge and reasoning abilities, rather than being influenced by format compliance issues, we employ GPT-4o to judge whether the final answers are equivalent.

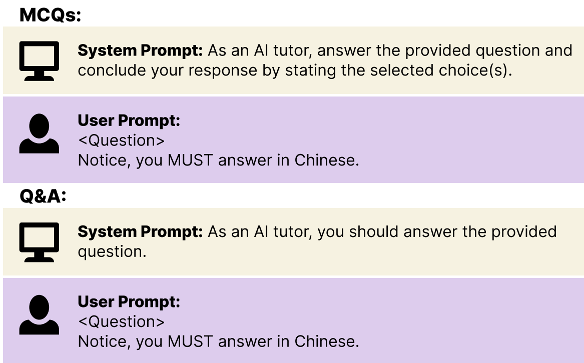


Figure 3: The prompt template designed for requesting models to answer questions in Chinese, where the <Question> is sourced from MMSciBench.

3.3 Prompt Design

We use prompts customized for different question types to evaluate the models in a zero-shot setting. For each question type, we apply the same specific prompt template across all models, avoiding model-specific prompt engineering that might explicitly guide reasoning or impose tailored requirements. The prompt template is illustrated in Fig. 3. To assess the models' intrinsic scientific abilities, the prompts used in the evaluation do not include additional key knowledge points or supplementary information from the dataset, although such infor-

Models	Math	Physics	Overall
Llama-3.2-90B-Vision-Instruct	16.69%	36.96%	31.19%
Gemini 1.5 Pro 002	56.74%	66.56%	63.77%
Claude 3.5 Sonnet	37.38%	60.54%	53.95%
GPT-4o	35.97%	56.89%	50.94%
Qwen2.5-Math-72B-Instruct	57.39%*	–	–
DeepSeekMath-7B-Instruct	21.86%*	–	–

Table 3: Accuracies of models across different subjects. Values marked with * indicate accuracies reported only on text-only questions, as the corresponding models are not multimodal.

mation could be incorporated in future research for other purposes. Since the dataset is in Chinese, we instruct the models to provide their answers in Chinese to ensure consistency with the dataset's language.

For the LLM-as-a-judge evaluation (Gu et al., 2024; Chen et al., 2024; Raju et al., 2024), we sample 180 instances of evaluated data and iteratively refined the judging prompts by manually verifying the accuracy of the judgments. This refinement process resulted in a judgment accuracy of 97.22%. Detailed prompts are provided in Sec. A in the appendix.

4 Results

4.1 Model Performance

Overall and Subject-wise Performance Table 3 presents the overall and subject-specific accuracies of the four LVLMs on the full MMSciBench dataset, along with the accuracies of the two math-specific LLMs on the text-only math subset. Gemini 1.5 Pro 002 achieves the highest overall accuracy at **63.77%**, significantly outperforming the other LVLMs in the evaluation. It consistently surpasses all competitors across each of the examined subjects, highlighting the substantial challenge posed by the benchmark, even for the most advanced LVLMs. Among the remaining LVLMs, Claude 3.5 Sonnet ranks second overall with an accuracy of **53.95%**, outperforming GPT-4o (**50.94%**) specifically in physics. In contrast, Llama-3.2-90B-Vision-Instruct lags far behind, recording the lowest overall accuracy of **31.19%**. For the two math-specific LLMs, Qwen2.5-Math-72B-Instruct demonstrates notable performance with an accuracy of **57.39%** on text-only math questions, while DeepSeekMath-7B-Instruct significantly underperforms, achieving only **21.86%**. This discrepancy is expected, given the difference in model sizes. Another noteworthy observation is the variation in performance across subjects, with models consistently performing better in physics. This finding will be analyzed further in Sec. 4.3.

Performance on Different Questions Types Table 4 reflects the performance of models on MCQs and Q&A questions in different subjects and the whole dataset, as well as the theoretical random-guess baselines. The random-guess baselines of MCQs are calculated based on the approximation that all MCQs in MMSciBench are 4-choice questions, as over 99% of MCQs in MMSciBench have 4 choices (see Table 7 in the appendix for detailed statistics). For single-choice questions, the random-guess accuracy is 1/4, as only one option is correct. For multiple-choice questions, where valid subsets include combinations of more than one choice, the random-guess accuracy is $1/(C_4^2 + C_4^3 + C_4^4) = 1/11$. For indeterminate-choice questions, where any non-empty subset of choices is valid, the random-guess accuracy is $1/2^4 = 1/16$. These probabilities were weighted to compute random-guess baselines of MCQs.

While the raw accuracies suggest that models generally perform better on MCQs than on Q&A questions, subtracting the baseline accuracies from their MCQ results reveals smaller yet positive gaps. This indicates that the provided answer choices in MCQs may assist the models by narrowing the possible answer space, making these questions easier to answer correctly compared to Q&A questions. Interestingly, this pattern does not hold true for math, where the MCQ advantage disappears after accounting for the baseline. In fact, some models seem to struggle more with MCQs than with Q&A questions in this subject. This suggests that the provided choices in math MCQs might mislead the models, making these questions more challenging.

4.2 Key Knowledge Point-Based Analysis

To better understand where different models excel or struggle within scientific domains—and to identify inherently challenging key knowledge points—all models' performances were analyzed across the taxonomy of first- and second-level key knowledge points, i.e., *Domain* and *Module* levels (see Fig. 4). This analysis reveals that, while models generally maintain consistent relative rankings across entire subjects, their strengths can vary significantly at the subfield level. For instance, although Gemini 1.5 Pro 002 often leads overall, it

Models	Math		Physics		Overall	
	MCQs	Q&A	MCQs	Q&A	MCQs	Q&A
Llama-3.2-90B-Vision-Instruct	25.39% 1.52%	3.88%	41.49% 21.48%	12.42%	37.96% 17.1%	8.08%
Gemini 1.5 Pro 002	63.16% 39.29%	47.29%	70.41% 50.40%	45.69%	68.82% 47.96%	46.50%
Claude 3.5 Sonnet	48.03% 24.16%	21.71%	65.35% 45.34%	34.47%	61.55% 40.69%	27.98%
GPT-4o	44.47% 20.60%	23.45%	61.17% 41.16%	33.67%	57.51% 36.65%	28.47%
Qwen2.5-Math-72B-Instruct	66.80%* 41.80%*	42.63%*	–	–	–	–
DeepSeekMath-7B-Instruct	32.40%* 7.40%*	5.33%*	–	–	–	–
Theoretical Random Baseline	23.87% 25.00%*	0 0*	20.01% –	0 –	20.86% –	0 –

Table 4: Accuracies of models across different question types, with underscored values indicating the accuracy improvement over the theoretical accuracy of random guess for MCQs. Values marked with * indicate accuracies on text-only math subsets for specialized math models.

falls behind Claude 3.5 Sonnet and GPT-4o in the subfield of "Electrodynamics - Magnetic Field". Additionally, certain subfields prove universally challenging, e.g., "Electrodynamics - Electromagnetic Induction and Its Applications" in physics, as well as "Geometry and Algebra – Geometry and Algebra" and "Functions – Preliminary Knowledge" in mathematics. These findings highlight both the nuanced capabilities and the current limitations of state-of-the-art models in addressing scientific knowledge.

4.3 Visual Understanding

MMSciBench includes both text-only and text-image paired questions. To evaluate the impact of visual input, we assess models on both types of questions, as shown in Table 5. Notably, all LVLMs perform worse on tasks involving both textual and visual elements compared to those relying solely on text. This highlights that bridging the gap between text comprehension and text-image co-reasoning remains a significant challenge for current LVLMs. Furthermore, the higher proportion of text-only questions in physics partially explains why models perform better on physics questions compared to math questions, as observed in Table 3.

4.4 The Effect of Chain-of-Thought in Reasoning

To evaluate the full scientific potential of the models, we design a suite of prompts to instruct them to answer step-by-step in Chinese, as detailed in Sec. A.2 in the appendix. As shown in Table 6, step-by-step prompting improves the accuracies of Llama-3.2-90B-Vision-Instruct and DeepSeekMath-7B-Instruct compared to their results in Table 3. However, the accuracy of Qwen2.5-Math-72B-Instruct decreases, while the performance of the other models remains unchanged.

This observation suggests that explicitly prompting certain models to use chain-of-thought reasoning can enhance their performance, and that different models exhibit varying degrees of alignment or readiness in this regard. Notably, Gemini 1.5 Pro 002, Claude 3.5 Sonnet, GPT-4o, and Qwen2.5-Math-72B-Instruct are more capable of generating effective reasoning steps without explicit prompting, whereas other models show more significant improvements when guided explicitly.

Considering that models typically have access to richer English training resources, we conducted additional experiments by prompting them to answer

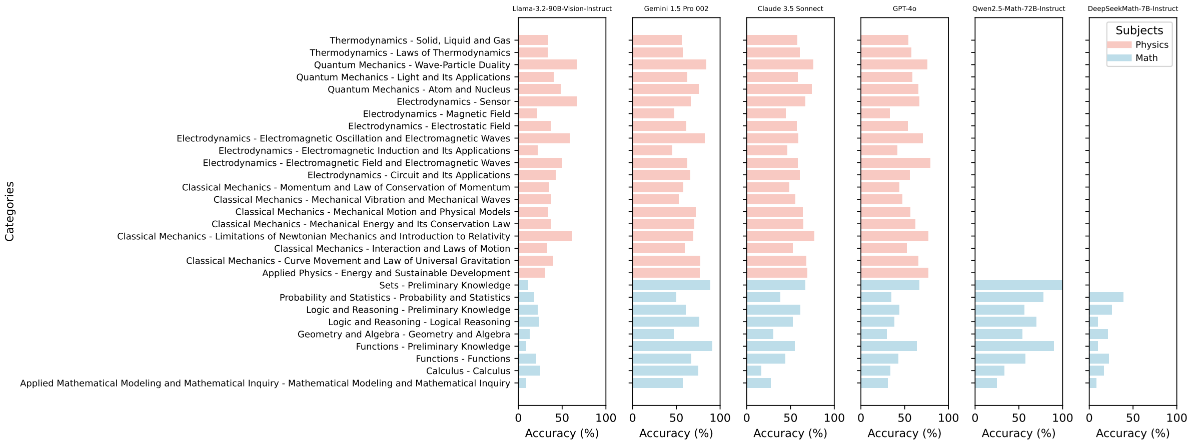


Figure 4: Accuracies of models across different key knowledge points.

Models	Math		Physics		Overall	
	Text	T&I	Text	T&I	Text	T&I
Llama-3.2-90B-Vision-Instruct	19.54%	11.60%	42.83%	16.34%	37.07%	14.48%
Gemini 1.5 Pro 002	69.60%	33.70%	74.40%	39.01%	73.21%	36.93%
Claude 3.5 Sonnet	44.57%	24.51%	67.75%	35.21%	62.02%	31.02%
GPT-4o	44.69%	20.35%	64.10%	31.55%	59.31%	27.16%
Qwen2.5-Math-72B-Instruct	57.39%	–	–	–	–	–
DeepSeekMath-7B-Instruct	21.86%	–	–	–	–	–

Table 5: Accuracies of models on text-only (**Text**) and text-image paired (**T&I**) questions across different subjects.

step-by-step in English to further explore their scientific capabilities. The corresponding prompts are detailed in Sec. A.2 of the appendix. As shown in Table 6, the results indicate that all models, except Gemini 1.5 Pro 002, benefit from this instruction. This underscores the effectiveness of explicit chain-of-thought prompting and its importance in accurately assessing models' capabilities. The differing behavior of Gemini 1.5 Pro 002 may suggest that its performance relies on the compatibility between the language of the questions and the language of the answers.

	Models	Math	Physics	Overall
in Chinese	**Llama-3.2-90B-Vision-Instruct**	19.12%	38.86%	33.24%
	Gemini 1.5 Pro 002	56.90%	66.28%	63.61%
	Claude 3.5 Sonnet	36.83%	61.42%	54.42%
	GPT-4o	35.74%	56.86%	50.85%
	Qwen2.5-Math-72B-Instruct	55.68%*	–	–
	DeepSeekMath-7B-Instruct	23.32%*	–	–
in English	**Llama-3.2-90B-Vision-Instruct**	22.41%	44.20%	38.00%
	Gemini 1.5 Pro 002	55.17%	65.07%	62.25%
	Claude 3.5 Sonnet	40.67%	61.26%	55.40%
	GPT-4o	37.23%	59.08%	52.86%
	Qwen2.5-Math-72B-Instruct	55.31%*	–	–
	DeepSeekMath-7B-Instruct	23.69%*	–	–

Table 6: Accuracies of models asked to provide step-by-step answers in Chinese and English. Values marked with * indicate accuracies on text-only math questions for the corresponding specialized math models.

5 Related Work

Scientific Benchmarks Scientific benchmarks are essential tools for evaluating the capabilities of language models in understanding and reasoning about complex scientific concepts, encompassing a wide range of disciplines, from general science to domain-specific areas like mathematics and physics. General scientific benchmarks, such as MSVEC (Evans et al., 2023) and SciOL (Tarsi et al., 2024), have been developed to assess various aspects of language models' abilities in specific scientific domains, including claim verification, figure retrieval, and multimodal information comprehension. However, the increasing complexity of language models necessitates more specialized benchmarks to evaluate their performance in specific scientific domains.

In mathematics, benchmarks like TRIGO (Xiong et al., 2023), DrawEduMath (Baral et al., 2025), and DMath (Kim et al., 2023) have been developed to assess AI models on targeted mathematical tasks. TRIGO focuses on formal mathematical proof reduction, evaluating models' abilities to understand and manipulate complex mathematical expressions. DrawEduMath is designed to assess models' proficiency in solving visual math problems, where both image and textual inputs are required to extract and process mathematical information. DMath, on the other hand, evaluates models on a diverse set of math word problems, testing their natural language understanding alongside mathematical reasoning. Similarly, in physics, datasets such as GRASP (Jassim et al., 2023) have been introduced to assess models' understanding of "Intuitive Physics" principles, including object permanence and continuity.

Additionally, benchmarks like GAOKAO-Bench (Zhang et al., 2023), GAOKAO-MM (Zong and Qiu, 2024), OlympiadBench (He et al., 2024), and SceMQA (Liang et al., 2024) span multiple scientific domains, including mathematics, physics, chemistry, and biology. These benchmarks focus on high-school, Olympiad, and pre-college levels, offering comprehensive evaluations of AI models' scientific reasoning capabilities across key disciplines.

Benchmarks for LVLMs Benchmarks for LVLMs have been developed to evaluate their performance across various tasks, including visual question answering, image captioning, and multimodal reasoning. These benchmarks typically consist of datasets with image-text pairs accompanied by corresponding questions or instructions, assessing the ability of LVLMs to generate accurate and relevant responses. For example, the VALSE benchmark (Parcalabescu et al., 2021) focuses on evaluating the visio-linguistic grounding capabilities of pretrained VLMs on specific linguistic phenomena. Other benchmarks, such as VisIT-Bench (Bitton et al., 2023), WinoGAViL (Bitton et al., 2022), and those designed for zero-shot visual reasoning (Nagar et al., 2024; Xu et al., 2024), are aimed at assessing the ability of LVLMs to reason about visual scenes and answer questions that require minimal world knowledge. These benchmarks often analyze the impact of conveying scene information either as visual embeddings or as purely textual scene descriptions to the underlying LLM of the LVLM.

To address the scarcity of scientific benchmarks specifically designed for the high school level—supporting both text-only and multimodal reasoning—we introduce MMSciBench. As detailed in Table 1, this dataset achieves a balanced trade-off between size and comprehensiveness, enabling efficient evaluation while offering a diverse selection of challenging high-school-level scientific problems. Additionally, MMSciBench prioritizes quality, with a significant portion of problems including detailed solution explanations and a three-level taxonomy of key knowledge points, facilitating fine-grained analysis of AI model performance.

6 Conclusion

This paper introduces MMSciBench, a benchmark designed to evaluate the scientific capabilities of both unimodal and multimodal language models. MMSciBench consists of a collection of high school-level MCQs and Q&A questions in mathematics and physics, with a subset of the questions incorporating images. The benchmark organizes its questions into a three-level taxonomy, ensuring comprehensive coverage of key knowledge points in both subjects. Our evaluation of four advanced LVLMs and two specialized math LLMs on MMSciBench demonstrates that current models still have significant room for improvement in scientific problem-solving. The analysis highlights that the inclusion of visual elements in questions presents a substantial challenge for model performance, emphasizing the complexity of integrating textual and visual reasoning. This work contributes to the ongoing development of robust benchmarks aimed at evaluating the evolving capabilities of language models, particularly in the domain of scientific reasoning.

Limitations

Despite the advances presented in MMSciBench, several limitations warrant discussion and open avenues for future research.

1. **Domain and Content Scope:** MMSciBench is focused on high-school level mathematics and physics, a scope chosen for its educational relevance and well-defined problem sets. However, this focus also limits the benchmark's applicability to broader scientific domains. While the curated questions capture es-

sential concepts, they do not encompass other fields such as chemistry, biology, or advanced scientific topics. Additionally, the dataset's reliance on K–12 educational standards may introduce biases that do not reflect the diverse challenges encountered in higher-level or interdisciplinary scientific reasoning.

2. **Evaluation Metrics and Reasoning Transparency:** The evaluation framework is centered on final answer accuracy, a metric that, while objective, does not capture the nuances of intermediate reasoning steps or the quality of explanations generated by models. By discounting partial correctness or the reasoning process, the assessment may obscure important differences in how models arrive at their answers. Future iterations of the benchmark may benefit from incorporating multi-faceted evaluation criteria that assess both the correctness of conclusions and the soundness of the reasoning process.

3. **Language and Cultural Considerations:** MMSciBench is primarily composed in Chinese, with some experiments extended to English. Models predominantly trained on English data may therefore be disadvantaged, and cultural or linguistic biases could affect performance. Future work should consider expanding the benchmark to include a more balanced representation of languages and educational contexts.

4. **Dataset Size and Filtering Practices:** While MMSciBench comprises **4,482** question–solution pairs, the dataset size is modest relative to some large-scale benchmarks. The strict filtering criteria (e.g., including only questions with a human-annotated hardness score ≥ 0.7) may also limit the diversity of problem difficulties, potentially excluding edge cases that could be valuable for assessing nuanced reasoning. Enlarging the dataset and diversifying the difficulty distribution would further strengthen the benchmark's comprehensiveness.

References

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

AI Anthropic. 2024. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*, 2.

Sami Baral, Li Lucy, Ryan Knight, Alice Ng, Luca Soldaini, Neil T Heffernan, and Kyle Lo. 2025. Drawedumath: Evaluating vision language models with expert-annotated students' hand-drawn math images. *arXiv preprint arXiv:2501.14877*.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. 2023. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*.

Yonatan Bitton, Nitzan Bitton Guetta, Ron Yosef, Yuval Elovici, Mohit Bansal, Gabriel Stanovsky, and Roy Schwartz. 2022. Winogavil: Gamified association benchmark to challenge vision-and-language models. *Advances in Neural Information Processing Systems*, 35:26549–26564.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Michael Evans, Dominik Soós, Ethan Landers, and Jian Wu. 2023. Msvec: A multidomain testing dataset for scientific claim verification. In *Proceedings of the Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pages 504–509.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

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

Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni. 2023. Grasp: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. *arXiv preprint arXiv:2311.09048*.

Jiwoo Kim, Youngbin Kim, Ilwoong Baek, JinYeong Bak, and Jongwuk Lee. 2023. It ain't over: A multi-aspect diverse math word problem dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14984–15011.

Zhenwen Liang, Kehan Guo, Gang Liu, Taicheng Guo, Yujun Zhou, Tianyu Yang, Jiajun Jiao, Renjie Pi, Jipeng Zhang, and Xiangliang Zhang. 2024. Scemqa: A scientific college entrance level multimodal question answering benchmark. *arXiv preprint arXiv:2402.05138*.

Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, Aixin Sun, Hany Awadalla, et al. 2024. Sciagent: Tool-augmented language models for scientific reasoning. *arXiv preprint arXiv:2402.11451*.

Aishik Nagar, Shantanu Jaiswal, and Cheston Tan. 2024. Zero-shot visual reasoning by vision-language models: Benchmarking and analysis. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2021. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*.

Ravi Raju, Swayambhoo Jain, Bo Li, Jonathan Li, and Urmish Thakker. 2024. Constructing domain-specific evaluation sets for llm-as-a-judge. *arXiv preprint arXiv:2408.08808*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

Henry W Sprueill, Carl Edwards, Mariefel V Olarte, Udishnu Sanyal, Heng Ji, and Sutanay Choudhury. 2023. Monte carlo thought search: Large language model querying for complex scientific reasoning in catalyst design. *arXiv preprint arXiv:2310.14420*.

Tim Tarsi, Heike Adel, Jan Hendrik Metzen, Dan Zhang, Matteo Finco, and Annemarie Friedrich. 2024. Sciol and mulms-img: Introducing a large-scale multimodal scientific dataset and models for image-text tasks in the scientific domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4560–4571.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Daniel Truhn, Jorge S Reis-Filho, and Jakob Nikolas Kather. 2023. Large language models should be used as scientific reasoning engines, not knowledge databases. *Nature medicine*, 29(12):2983–2984.

Jing Xiong, Jianhao Shen, Ye Yuan, Haiming Wang, Yichun Yin, Zhengying Liu, Lin Li, Zhijiang Guo, Qingxing Cao, Yinya Huang, et al. 2023. Trigo: Benchmarking formal mathematical proof reduction for generative language models. *arXiv preprint arXiv:2310.10180*.

Zhenlin Xu, Yi Zhu, Siqi Deng, Abhay Mittal, Yanbei Chen, Manchen Wang, Paolo Favaro, Joseph Tighe, and Davide Modolo. 2024. Benchmarking zero-shot recognition with vision-language models: Challenges on granularity and specificity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1827–1836.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*.

Yi Zong and Xipeng Qiu. 2024. Gaokao-mm: A chinese human-level benchmark for multimodal models evaluation. *arXiv preprint arXiv:2402.15745*.

A Prompts

In this section, we present the prompts used in our work.

A.1 The Prompt for Question Categorization

Fig. 5 presents the prompt designed for categorizing MMSciBench questions into specific categories using GPT-4o. The category sets for each subject are derived from a Chinese high school key knowledge point taxonomy.

User Prompt:
你是一个分类助手，用于将提供的习题题目归类到以下类别之一：
<Categories>
请先进行分析，然后在最后以以下格式提供你的分类结果：
分类结果: <类别名称>
请确保你的分类结果严格遵守上述格式，并且不要修改类别名称。
<Question>

Figure 5: The prompt template is designed to use GPT-4 as a classifier, categorizing each question into a three-level hierarchy. <Categories> represents the predefined set of categories for the target subject.

A.2 Prompt Templates for the Effect of Chain-of-Thought in Reasoning

Fig. 6 and Fig. 7 are prompts templates that ask models to think step by step in Chinese and English, respectively.

MCQs:

System Prompt: As an AI tutor, answer the provided question and conclude your response by stating the selected choice(s).

User Prompt:
<Question>
Notice, you MUST answer in Chinese. Let's solve it step by step.

Q&A:

System Prompt: As an AI tutor, you should answer the provided question.

User Prompt:
<Question>
Notice, you MUST answer in Chinese. Let's solve it step by step.

Figure 6: The prompt template is designed for requesting models to answer questions in Chinese step by step, where the <Question> is sourced from MMSciBench.

A.3 The Prompt Template for Using GPT-4o as a Judge

Fig. 8 (with its English translation in Fig. 9) illustrates the prompt used to instruct GPT-4o to evaluate whether a "student solution"—that is, the model's response being assessed—is correct or incorrect compared to the standard solution in MMSciBench. For MCQs, only the model's answer and

MCQs:

System Prompt: As an AI tutor, answer the provided question and conclude your response by stating the selected choice(s).

User Prompt:
<Question>
Notice, you MUST answer in English. Let's solve it step by step.

Q&A:

System Prompt: As an AI tutor, you should answer the provided question.

User Prompt:
<Question>
Notice, you MUST answer in English. Let's solve it step by step.

Figure 7: The prompt template is designed for requesting models to answer questions in English step by step, where the <Question> is sourced from MMSciBench.

the standard solution are provided, omitting the actual questions. This approach is sufficient because the evaluation solely involves comparing whether the selected choices match the standard answer, eliminating the need to understand the question's context. In contrast, for Q&A questions, GPT-4o is provided with the question, the standard solution, and the model's answer. This comprehensive context enables accurate semantic understanding and a thorough comparison between the two responses. The prompt for Q&A questions have been iteratively refined and enhanced to improve GPT-4o's judgment, particularly in cases where misjudgments are likely. This refinement process involves sampling a subset of evaluated responses and manually diagnosing the reasons for any misjudgments, thereby continually improving the evaluation accuracy.

MCQs:

System Prompt: 你是一个助教助手，负责判断学生答案的选择是否与标准答案一致。

User Prompt:
标准答案：<Standard Solution>
学生答案：<Student Solution>
标准答案只包含选项，而学生答案可能包含思考过程或解析。
你需要从学生答案中提取具体选择项，并与标准答案进行比对。
如果提取出的学生选择与标准答案完全一致，则回答"正确"，否则回答"错误"。
判断结果以"正确"或"错误"的形式回答，不要提供任何其他信息。

Q&A:

System Prompt: 你是一个助教助手，负责判断学生答案的结论是否与标准答案的结论表达同样的意思。

User Prompt:
请根据以下题目信息和提供的标准答案判断学生答案是否正确:
问题：<Question>
标准答案：<Standard Solution>
学生答案：<Student Solution>
请根据以上题目信息和标准答案，仅根据学生答案的最终结论或答案判断其是否正确，忽略过程的正确性。

注意事项：
1. 检查题目是否包含多个子问题：
- 如果包含多个子问题，请逐一判断每个子问题的答案是否正确。只有当所有子问题的最终答案都正确时，整体答案才被视为正确。
- 如果不包含子问题，则仅根据学生答案的最终结论或答案进行判断。
2. 即使学生答案的表达方式与标准答案不同，只要最终结论或答案的意思相同，也应视为正确。可能的情况包括但不限于：
- 学生答案和标准答案使用的语言不同，但意思相同。
- 学生答案中的公式经过化简和变形后与标准答案的公式相同。
- 学生答案采用了不同的表述形式，但语义相同。
3. 请解释和分析学生答案与标准答案的最终结论或答案的异同之处。
4. 如果由于学生答案不完整或缺失，或者学生答案没有按照题目要求给出结论，导致无法判断是否正确的，就判断为错误。
5. 如果题目包含子问题，请为每个子问题提供"子问题X判断结果：正确"或"子问题X判断结果：错误"的形式。
6. 最终判断结果应以"判断结果：正确"或"判断结果：错误"的形式给出。

请按照以下格式回复：

分析：
[在此处填写详细的分析内容]

[如果有子问题，则添加以下部分]
子问题判断结果：
子问题1判断结果：正确/错误
子问题2判断结果：正确/错误
...
子问题N判断结果：正确/错误

判断结果：正确/错误

Figure 8: The prompt template designed for using GPT-4o as a judge, where the <Question> and <Standard Solution> is sourced from MMSciBench, while <Student Solution> is the solution provided by the tested model.

MCQs:

System Prompt: You are a teaching assistant responsible for determining whether the choices of students' solution match the standard solution.

User Prompt:
Standard Solution: <Standard Solution>
Student Solution: <Student Solution>

Standard solution only contains the choices, while student solution may include reasoning or explanations.
You need to extract the specific choices from the student solution and compare them with the standard solution.

If the extracted student choices match the standard solution exactly, respond with "Correct"; otherwise, respond with "Incorrect."
The judgment should be provided in the form of "Correct" or "Incorrect" only, without any additional information.

Q&A:

System Prompt: You are a teaching assistant responsible for determining whether the conclusion of the student solution expresses the same meaning as the conclusion of the standard solution.

User Prompt:
Please determine whether the student solution is correct based on the following question information and the provided standard solution:

Question: <Question>
Standard Solution: <Standard Solution>
Student Solution: <Student Solution>

Make your judgment based solely on the final conclusion or answer provided in the student solution, ignoring the correctness of the process.

Notes:
1. Check whether the question contains multiple sub-questions:
- If it contains multiple sub-questions, evaluate each sub-question individually to determine whether its answer is correct. Only when the final answers to all sub-questions are correct is the overall answer considered correct.
- If there are no sub-questions, judge based solely on the final conclusion or answer in the student solution.
2. Even if the student's expression differs from the standard solution, as long as the final conclusion or answer conveys the same meaning, it should be considered correct. Possible cases include but are not limited to:
- The language used in the student solution differs from the standard solution, but the meaning is the same.
- The formula in the student solution simplifies or transforms into the same formula as the standard solution.
- The student solution uses a different expression, but the semantics are identical.
3. Explain and analyze the similarities and differences between the final conclusion or answer of the student solution and the standard solution.
4. If the student solution is incomplete, missing, or does not provide a conclusion as required by the question, making it impossible to determine correctness, the judgment should be "Incorrect."
5. If the question contains sub-questions, provide results for each sub-question in the format of "Sub-question X result: Correct" or "Sub-question X result: Incorrect."
6. The final judgment should be given in the format: "Judgment Result: Correct" or "Judgment Result: Incorrect."

Please folllow the following response format:

Analysis:
[Provide detailed analysis here]

[If there are sub-questions, include the following section]
Sub-question Results:
Sub-question 1 result: Correct/Incorrect
Sub-question 2 result: Correct/Incorrect
...
Sub-question N result: Correct/Incorrect

Judgment Result: Correct/Incorrect

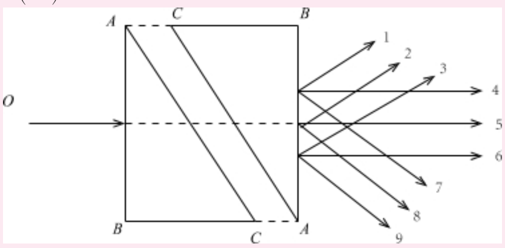
Figure 9: The English translation of the prompt template shown in Fig. 8.

B Data Examples

In this section, we present examples from MM-SciBench, including a physics MCQ (Fig. 10 and the corresponding English translation in Fig. 1), a physics Q&A question (Fig. 15 and the corresponding English translation in Fig. 16), a math MCQ (Fig. 11 and the corresponding English translation in Fig. 12), and a math Q&A question (Fig. 13 and the corresponding English translation in Fig. 14). Each example is accompanied by its standard solution and explanation.

Question & Standard Solution

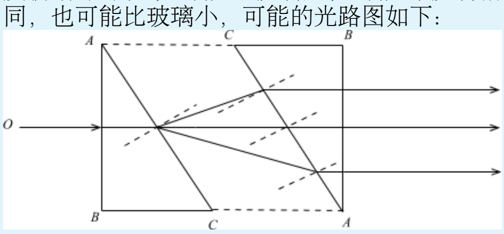
Question
问题（单选）：如图所示，两块同样的玻璃直角三棱镜ABC，两者的AC面是平行放置的，在它们之间是均匀的未知透明介质。一束单色细光O垂直于AB面入射，在图示的出射光线中（ ）。



选项:
A. 1、2、3（彼此平行）中的任一条都有可能
B. 4、5、6（彼此平行）中的任一条都有可能
C. 7、8、9（彼此平行）中的任一条都有可能
D. 只能是4、6中的某一条
Standard Solution: B

Explanation

本题主要考查三棱镜问题。
选项分析：据题述，两个直角三棱镜之间的介质折射率未知，可能比玻璃大，可能与玻璃相同，也可能比玻璃小，可能的光路图如下：

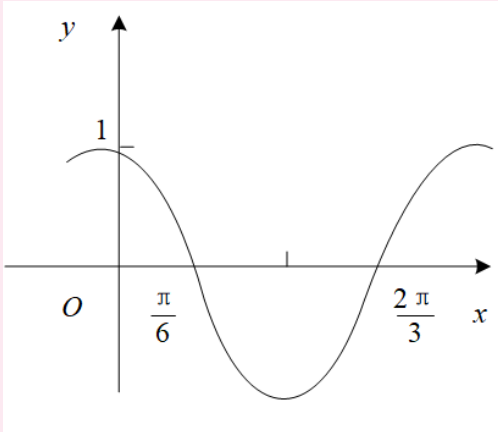


故B项正确，ACD项错误。
综上所述，本题正确答案为B。

Figure 10: An example of a physics MCQ.

Question & Standard Solution

Question
问题（多选）：下图是函数$y = \sin(\omega x + \varphi)$的部分图象，则$\sin(\omega x + \varphi) =$（）。



选项:
A. $\sin(x + \frac{\pi}{3})$
B. $\sin(\frac{\pi}{3} - 2x)$
C. $\cos(2x + \frac{\pi}{6})$
D. $\cos(\frac{5\pi}{6} - 2x)$
Standard Solution: B, C

Explanation

本题主要考查三角函数。
由题图可知，

$$\frac{T}{2} = \frac{2}{3}\pi - \frac{\pi}{6} = \frac{\pi}{2},$$

所以

$$T = \frac{2\pi}{|\omega|} = \pi,$$

所以$|\omega| = 2$。
当$\omega = 2$时，由函数图象过点$(\frac{\pi}{6}, 0)$，$(\frac{2\pi}{3}, 0)$，且$f(0) > 0$，得

$$\varphi = \frac{2\pi}{3} + 2k\pi \quad (k \in \mathbb{Z}),$$

所以

$$y = \sin\left(2x + \frac{2\pi}{3}\right) = -\cos\left(\frac{5\pi}{6} - 2x\right),$$

同理，当$\omega = -2$时，

$$\varphi = \frac{\pi}{3} + 2k\pi \quad (k \in \mathbb{Z}),$$

所以

$$y = \sin\left(-2x + \frac{\pi}{3}\right) = \cos\left(2x + \frac{\pi}{6}\right)$$

故本题正确答案为BC。

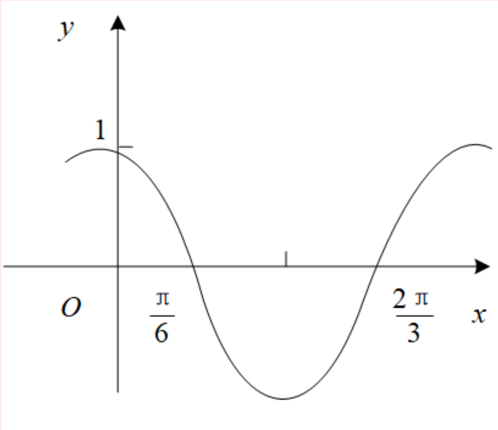
Figure 11: An example of a math MCQ.

C The Distribution of Choices of MCQs

Table 7 shows that over 99% of MCQs in MM-SciBench have 4 choices

Question & Standard Solution

Question
Question (Multiple Choice): The figure below shows a part of the graph of the function $y = \sin(\omega x + \varphi)$. Determine $\sin(\omega x + \varphi) = ()$.



Options:
A. $\sin\left(x + \frac{\pi}{3}\right)$
B. $\sin\left(\frac{\pi}{3} - 2x\right)$
C. $\cos\left(2x + \frac{\pi}{6}\right)$
D. $\cos\left(\frac{5\pi}{6} - 2x\right)$
Standard Solution: B, C

Explanation

This question primarily assesses trigonometric functions.
From the figure, we know that

$$\frac{T}{2} = \frac{2}{3}\pi - \frac{\pi}{6} = \frac{\pi}{2},$$

therefore

$$T = \frac{2\pi}{|\omega|} = \pi,$$

so $|\omega| = 2$.
When $\omega = 2$, since the graph passes through the points $\left(\frac{\pi}{6}, 0\right)$ and $\left(\frac{2\pi}{3}, 0\right)$, and $f(0) > 0$, we have

$$\varphi = \frac{2\pi}{3} + 2k\pi \quad (k \in \mathbb{Z}),$$

thus

$$y = \sin\left(2x + \frac{2\pi}{3}\right) = -\cos\left(\frac{5\pi}{6} - 2x\right),$$

similarly, when $\omega = -2$,

$$\varphi = \frac{\pi}{3} + 2k\pi \quad (k \in \mathbb{Z}),$$

so

$$y = \sin\left(-2x + \frac{\pi}{3}\right) = \cos\left(2x + \frac{\pi}{6}\right)$$

Therefore, the correct answer is BC.

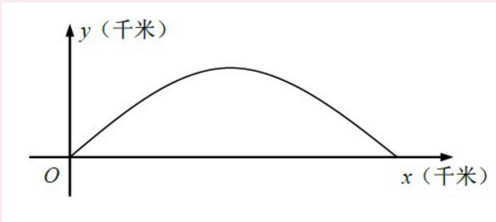
Figure 12: The English translation of the math MCQ example in Fig. 11.

Question & Standard Solution

Question
问题（解答）：如图，建立平面直角坐标系xOy，x轴在地平面上，y轴垂直于地平面，单位长度为1千米。某炮位于坐标原点。已知炮弹发射后的轨迹在方程

$$y = kx - \frac{1}{20}(1+k^2)x^2 (k > 0)$$

表示的曲线上，其中k与发射方向有关。炮的射程是指炮弹落地点的横坐标。（1）求炮的最大射程；（2）设在第一象限有一飞行物（忽略其大小），其飞行高度为3.2千米，试问它的横坐标a不超过多少时，炮弹可以击中它？请说明理由。



Standard Solution
（1）令$y = 0$，得$kx - \frac{1}{20}(1+k^2)x^2 = 0$，由实际意义和题设条件知$x > 0$，$k > 0$，故

$$x = \frac{20k}{1+k^2} = \frac{20}{k + \frac{1}{k}} \le \frac{20}{2} = 10,$$

当且仅当$k = 1$时取等号。所以炮的最大射程为10千米。
（2）因为$a > 0$，所以炮弹可击中目标
⇔存在$k > 0$，使$3.2 = ka - \frac{1}{20}(1+k^2)a^2$成立
⇔关于k的方程$a^2k^2 - 20ak + a^2 + 64 = 0$有正根
⇔判别式

$$\Delta = (-20a)^2 - 4a^2(a^2 + 64) \ge 0$$

⇔ $a \le 6$
此时，

$$k = \frac{20a + \sqrt{(-20a)^2 - 4a^2(a^2+64)}}{2a^2} > 0$$

（不考虑另一根）。所以当a不超过6千米时，可击中目标。

Explanation

本题主要考查函数与方程和基本不等式的应用等相关知识。（1）求炮的最大射程，即$y = 0$时的一个较大的根，因为含有参数k，所以需根据k的取值范围确定另外一个根的最大值，即为炮的最大射程。（2）炮弹能击中目标的含义为炮弹的飞行高度$y = 3.2$时有解。根据二次函数有正根，可得出a的取值范围。

Figure 13: An example of a math Q&A question.

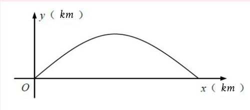
Question & Standard Solution

Question

Question (Q&A): As shown in the figure, set up a Cartesian coordinate system xOy, with the x-axis on the ground, the y-axis perpendicular to the ground, and the unit length is 1 kilometer. A cannon is located at the origin. It is known that the trajectory of the cannonball after firing is represented by the equation

$$y = kx - \frac{1}{20}(1+k^2)x^2(k>0)$$

where k is related to the firing direction. The cannon's range refers to the x-coordinate of the landing point of the cannonball. (1) Find the maximum range of the cannon; (2) Suppose there is a flying object in the first quadrant (ignoring its size) with a flight height of 3.2 kilometers. What is the maximum x-coordinate a such that the cannonball can hit it? Please explain your reasoning.



Standard Solution

(1) Set $y=0$, obtaining $kx - \frac{1}{20}(1+k^2)x^2 = 0$. From the actual meaning and problem conditions, we know $x>0$, $k>0$, thus

$$x = \frac{20k}{1+k^2} = \frac{20}{k+\frac{1}{k}} \leq \frac{20}{2} = 10,$$

equality holds if and only if $k=1$. Therefore, the maximum range of the cannon is 10 kilometers.
(2) Because $a>0$, the cannonball can hit the target
$\Leftrightarrow$ there exists $k>0$ such that $3.2 = ka - \frac{1}{20}(1+k^2)a^2$ holds
$\Leftrightarrow$ the equation $a^2k^2 - 20ak + a^2 + 64 = 0$ in terms of k has positive roots
$\Leftrightarrow$ the discriminant

$$\Delta = (-20a)^2 - 4a^2(a^2+64) \geq 0$$

$\Leftrightarrow a \leq 6$
At this time,

$$k = \frac{20a + \sqrt{(-20a)^2 - 4a^2(a^2+64)}}{2a^2} > 0$$

(Not considering the other root). Therefore, when a does not exceed 6 kilometers, the target can be hit.

Explanation

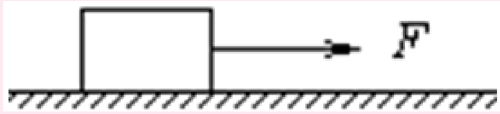
This question primarily tests the application of functions, equations, and basic inequalities. (1) To find the maximum range of the cannon, which is the larger root when $y=0$, because there is a parameter k, we need to determine the maximum value of the other root based on the range of k, which gives the cannon's maximum range. (2) The meaning of the cannonball being able to hit the target is that when the flight height $y=3.2$, there exists a solution. Based on the quadratic function having positive roots, we can derive the range of a.

Figure 14: The English translation of the math Q&A question example in Fig. 13.

Question & Standard Solution

Question

问题（解答）：如图所示，在光滑的水平面上，质量$m=5kg$的物体，在水平拉力$F=10N$的作用下，从静止开始运动，运动时间$t=3s$。求：（1）力F在$3s$内对物体所做的功；（2）力F在$3s$内对物体做功的平均功率；（3）在$3s$末，力F对物体做功的瞬时功率。



Standard Solution

（1）由牛顿第二定律可得：$F=ma$，$3s$内对物体的位移为$x=\frac{1}{2}at^2$，则力F在$3s$内对物体所做的功为$W=Fx$，联立可得：$W=90J$。
（2）力F在$3s$内对物体做功的平均功率为$\overline{P}=\frac{W}{t}=30W$。
（3）在$3s$末物体的速度大小为$v=at$，则在$3s$末，力F对物体做功的瞬时功率为$P=Fv$，联立可得：$P=60W$。

Explanation

本题主要考查牛顿第二定律和功率公式的选择与计算。
问题求解：
（1）由牛顿第二定律可算出运动的加速度，便可求出$3s$内对物体的位移，便能算出力F在$3s$内对物体所做的功。
（2）根据$\overline{P}=\frac{W}{t}$便可算出力F在$3s$内对物体做功的平均功率。
（3）先算出在$3s$末物体的速度大小，根据$P=Fv$便可算出在$3s$末，力F对物体做功的瞬时功率。

Figure 15: An example of a physics Q&A question.

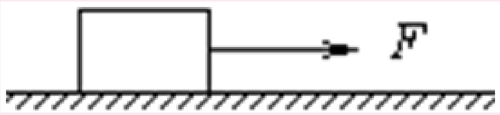
Subject	**Image**	**4 Choices**	**Other**	**Total**
Physics	✗	2230	27	2257
Physics	✓	448	2	450
Math	✗	500	0	500
Math	✓	260	0	260
Total		3438	29	3467

Table 7: Distribution of choice numbers in MCQs in MMSciBench by subject and image presence.

Question & Standard Solution

Question

Question (Q&A): As shown in the figure, on a smooth horizontal plane, a mass $m = 5kg$ object is acted upon by a horizontal force $F = 10N$ and starts moving from rest. The motion time is $t = 3s$. Find: (1) The work done by force F on the object within $3s$; (2) The average power of force F in doing work on the object within $3s$; (3) The instantaneous power of force F in doing work on the object at the end of $3s$.



Standard Solution

(1) From Newton's second law, $F = ma$. The displacement of the object within $3s$ is $x = \frac{1}{2}at^2$. Therefore, the work done by force F on the object within $3s$ is $W = Fx$. Solving these equations yields $W = 90J$.
(2) The average power of force F in doing work on the object within $3s$ is $\overline{P} = \frac{W}{t} = 30W$.
(3) At the end of $3s$, the velocity of the object is $v = at$. Therefore, the instantaneous power of force F in doing work on the object at the end of $3s$ is $P = Fv$. Solving these equations yields $P = 60W$.

Explanation

This problem primarily tests the application and calculation of Newton's second law and power formulas.
Problem Solving:
(1) Using Newton's second law, the acceleration of the motion can be calculated, which allows us to find the displacement of the object within 3 s. This displacement can then be used to calculate the work done by force F on the object within $3s$.
(2) Using $\overline{P} = \frac{W}{t}$, the average power of force F in doing work on the object within $3s$ can be calculated.
(3) First, calculate the velocity of the object at the end of $3s$. Then, using $P = Fv$, the instantaneous power of force F in doing work on the object at the end of $3s$ can be calculated.

Figure 16: The English translation of the physics Q&A question example in Fig. 15.