MMSciBench: Benchmarking Language Models on Chinese 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., 2023b; 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 five 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.

Our evaluation reveals significant limitations in current models' multimodal scientific reasoning

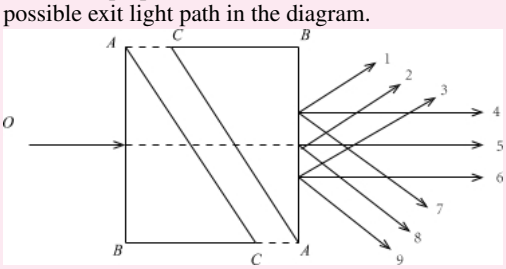
*Corresponding Authors.

[1] https://github.com/xinwuye/MMSciBench-code

[2] https://huggingface.co/datasets/XinwuYe/MMSciBench

Question & Standard Solution

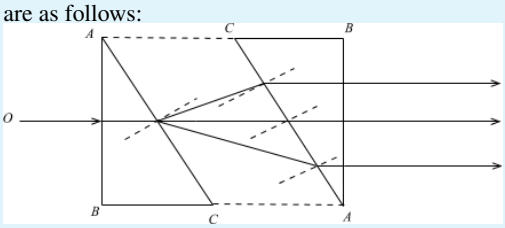
Question
Question (Single Choice): As shown in the figure, two identical right-angled glass prisms ABC are placed with their AC faces parallel to each other, and between them is a uniform unknown transparent medium. A monochromatic thin light beam O is incident perpendicular to the AB face. ( ) is the 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
Difficulty Level: 0.7
Domain: Quantum Mechanics
Module: Light and Its Applications
Chapter: Snell's Law
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. 11 in the appendix.

capabilities. Gemini 1.5 Pro 002 achieved the highest accuracy (**63.77%**), followed by Qwen2-VL-72B-Instruct (**56.11%**), 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 Gemini 1.5 Pro 002's 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 Gemini 1.5 Pro 002's **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 range of other metadata, including question type (MCQ/Q&A), modality (text-only/text-image), and subject (math/physics). The detailed curation process is described in Sec. A in the appendix. 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. The benchmark includes standardized problem prompts to ensure consistent model input. The benchmark also integrates a GPT-4o evaluator to assess answer correctness, focusing on scientific capability over format adherence. 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

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

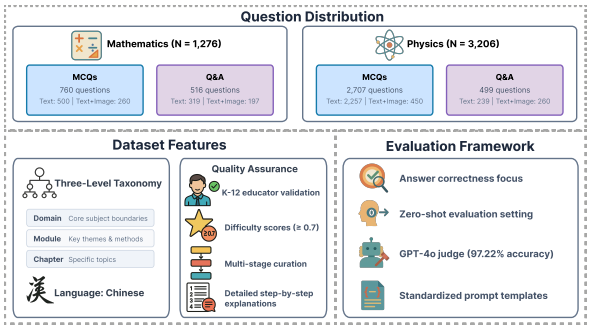


Figure 2: The overview of MMSciBench, describing the question distribution, dataset features, and the evaluation framework.

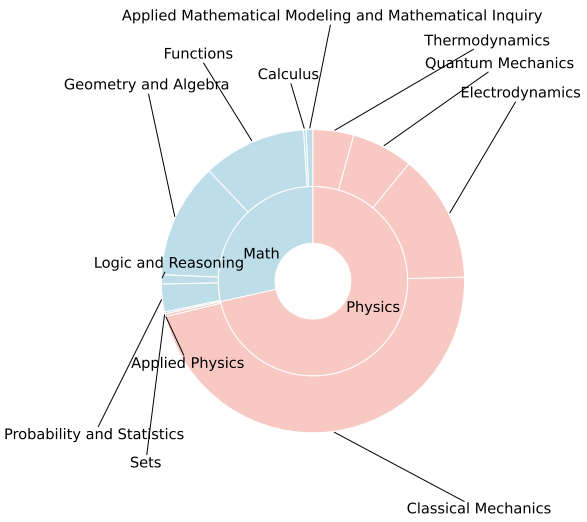


Figure 3: The distribution of data in MMSciBench according to the first-level key knowledge points for each subject.

employed GPT-4o to annotate each question according to a three-level subject-specific taxonomy, detailed in Sec. 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 that enable rigorous evaluation of models' scientific reasoning capabilities across diverse domains.

2.2 Dataset Description

Fig. 2 provides a visual overview of MMSciBench, detailing the distribution of questions in the dataset, dataset features, and the evaluation framework.

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. 11 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 3.

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.
- **Module**: Subdivisions within *Domains* that focus on key themes or methods. Examples include "Probability and Statistics" in mathematics and "Mechanical Motion and Physical

Benchmark	Subject(s)	Modality	Key Knowledge Pt.	Explanation	Language	Difficulty	Size
TRIGO (Xiong et al., 2023)	M	T	✗	✓	Lean	High School	11K
DMath (Kim et al., 2023)	M	T	✓	✓	EN&KR	Grade School	10K
GRASP (Jassim et al., 2023)	P	T&V	✓	✗	EN	Basic	2K
MSVEC (Evans et al., 2023)	P, O	T	✗	✓	EN	College	200
SciOL (Tarsi et al., 2024)	P, O	T&I	✗	✗	EN	College	18M
SciEval (Sun et al., 2024)	P, O	T	✓	Partial	EN	Multi-level	16K
SceMQA (Liang et al., 2024)	M, P, O	T&I	✓	✓	EN	Pre-College	1K
GAOKAO-Bench (Zhang et al., 2023b)	M, P, O	T	✗	✓	ZH	High School	3K
GAOKAO-MM (Zong and Qiu, 2024)	M, P, O	T, T&I	✗	✓	ZH	High School	650
SciBench (Wang et al.)	M, P, O	T, T&I	✓	✓	EN	College	869
MMSci (Li et al., 2024)	Multi-subj. (72)	T&I	✓	✓	EN	PhD-level	108K
M3Exam (Zhang et al., 2023a)	Multi-subj.	T, T&I	✓	(A)	Multi. (9 lang.)	K-12 levels	12K
SciFIBench (Roberts et al., 2024)	Multi-subj.	T&I	✓	✓	EN	Academic	2K
EXAMS-V (Das et al., 2024)	Multi-subj. (20)	T&I	✓	(A)	Multi. (11 lang.)	School Exams (4-12)	21K
OlympiadBench (He et al., 2024)	M, P	T, T&I	✓	✓	EN, ZH	Olympiad	8K
MMSciBench (Ours)	M, P	T, T&I	✓	✓	ZH	High School	4K

Table 1: Comparison of MMSciBench with existing benchmarks. For Subject(s), 'M' denotes mathematics, 'P' denotes physics, and 'O' denotes other subject(s). For Modality, 'T' denotes text-only data, 'T&I' denotes text-image data pairs, and 'T&V' denotes text-video data pairs. EN, ZH, KR, and Lean represent English, simplified Chinese, Korean, and the Lean theorem prover language, respectively. Multi. denotes multilingual. For Explanation, '(A)' indicates answers are provided.

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.

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 five state-of-the-art LVLMs: GPT-4o, Claude 3.5 Sonnet (Anthropic, 2024), Gemini 1.5 Pro 002 (Team et al., 2024), Llama-3.2-90B-Vision-Instruct, and Qwen2-VL-72B-Instruct (Wang et al., 2024a).

In addition, we evaluated several other state-of-the-art models. For models specifically designed for mathematical problem-solving, we included Qwen2.5-Math-72B-Instruct (Yang et al., 2024) and DeepSeekMath-7B-Instruct (Shao et al., 2024) on text-only mathematics questions. To further broaden our comparison, particularly on multi-modal reasoning which is a key focus of MMSciBench, we evaluated state-of-the-art reasoning models, including o1 (Jaech et al., 2024) and Claude 3.7 Sonnet (Anthropic, 2025), on the text-image math questions subset of our benchmark. 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. Given the full response of tested models, GPT-4o is then employed to compare the model's final answer directly against the ground truth to assess correctness. In existing studies, MCQs often require models to adhere to a specified output

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.

Q&A:

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

User Prompt:
<Question>
Notice, you MUST answer in Chinese.

Figure 4: The prompt template designed for requesting models to answer questions in Chinese, where the <Question> is sourced from MMSciBench.

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-VL-72B-Instruct	35.50%	64.32%	56.11%
Qwen2.5-Math-72B-Instruct	57.39%*	–	–
DeepSeekMath-7B-Instruct	21.86%*	–	–
o1	67.40%†	–	–
Claude 3.7 Sonnet	37.64%†	–	–

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. Values marked with † indicate accuracies reported only on the text-image questions.

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.

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. 4. 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 information 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 achieved a **97.22%** judgment accuracy, the agreement rate between GPT-4o's and human evaluations. The GPT-4o evaluator outputs a deterministic final judgment in a required format. Detailed prompts are provided in Sec. B in the appendix.

To further analyze the reliability of GPT-4o as a judge, we conducted a qualitative study of its accuracy on this task. In this study, we first identified all instances incorrectly judged by GPT-4o from our evaluation sample and randomly selected an additional **50** correctly-judged instances from the same sample for comparative analysis. We then classified the error patterns in the incorrectly judged cases and documented representative examples of both successful and failed judgments by GPT-4o. (Further examples are provided in Sec. E in the appendix). Our analysis revealed that GPT-4o generally excels at understanding semantic equivalence between a model's generated answer and the provided standard solution. However, occasional errors in its judging capabilities were observed. These errors primarily stemmed from two main causes: (1) misinterpretations of the standard of correctness (e.g., being too lenient on incomplete answers or overlooking errors in multi-part questions), and (2) mistakenly equating two distinct mathematical formulas as equivalent when they were not. It underscores that definitions of 'correctness' and the detailed comparison of complex mathematical expressions are sources of potential discrepancy.

4 Results

4.1 Model Performance

Overall and Subject-wise Performance Table 3 presents the overall and subject-specific accuracies of the five LVLMs on the full MMSciBench dataset, the accuracies of the two math-specific

LLMs on the text-only math subset, and the three reasoning models on the text-image 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, Qwen2-VL-72B-Instruct ranks second overall with an accuracy of **56.11%**, outperforming Claude 3.5 Sonnet (**53.95%**) and GPT-4o (**50.94%**). 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. Furthermore, on the text-image math questions subset, o1 achieves an accuracy of **67.40%**, outperforming Claude 3.7 Sonnet (**37.64%**). These results on a challenging multimodal subset further highlight the varying capabilities of different models in utilizing visual and textual information for mathematical reasoning. 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. Additionally, a breakdown of performance by question difficulty is provided in Sec. F of the appendix.

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.

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-VL-72B-Instruct	46.58% 22.71%	19.19%	71.07% 51.06%	27.66%	65.71% 44.85%	23.35%
Qwen2.5-Math-72B-Instruct	66.80%* 41.80%*	42.63%*	–	–	–	–
DeepSeekMath-7B-Instruct	32.40%* 7.40%*	5.33%*	–	–	–	–
o1	71.54%† 49.86%†	61.93%†	–	–	–	–
Claude 3.7 Sonnet	44.62%† 22.94%†	28.43%†	–	–	–	–
Theoretical Random Baseline	23.87% 25.00%* 21.68%†	0 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 subsets. Values marked with † indicate accuracies reported only on the text-image questions.

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 consistently 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 Taxonomy-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. 5). This analysis reveals that, while models generally maintain consistent relative rankings across entire subjects, their strengths can vary

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-VL-72B-Instruct	41.39%	24.95%	72.48%	35.63%	64.80%	31.45%
Qwen2.5-Math-72B-Instruct	57.39%	–	–	–	–	–
DeepSeekMath-7B-Instruct	21.86%	–	–	–	–	–
o1	–	67.40%	–	–	–	–
Claude 3.7 Sonnet	–	37.64%	–	–	–	–

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

	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-VL-72B-Instruct	30.72%	57.77%	50.07%
	Qwen2.5-Math-72B-Instruct	55.68%*	–	–
	DeepSeekMath-7B-Instruct	23.32%*	–	–
	o1	68.05%†	–	–
	Claude 3.7 Sonnet	39.61%†	–	–
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-VL-72B-Instruct	32.68%	60.79%	52.79%
	Qwen2.5-Math-72B-Instruct	55.31%*	–	–
	DeepSeekMath-7B-Instruct	23.69%*	–	–
	o1	68.49%†	–	–
	Claude 3.7 Sonnet	36.32%†	–	–

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. Values marked with † indicate accuracies reported only on the text-image math questions.

significantly at the subfield level. For instance, although Gemini 1.5 Pro 002 often leads among non-reasoning models, it 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. B.2 in the appendix. As shown in Table 6, step-by-step prompting improves the accuracies of Llama-3.2-90B-Vision-Instruct, DeepSeekMath-7B-Instruct, o1, and Claude 3.7 Sonnet compared to their results in Table 3. However, the accuracy of Qwen2.5-Math-72B-Instruct and Qwen2-VL-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, Qwen2.5-Math-72B-Instruct, and Qwen2-VL-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 step-by-step in English to further explore their scientific capabilities. The corresponding prompts are detailed in Sec. B.2 of the appendix. As shown in Table 6, the results indicate that models, except Gemini 1.5 Pro 002 and Claude 3.7 Sonnet, 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 the two models may suggest that their performance relies on the compatibility between the language of the questions and the language of the answers.

4.5 Error Analysis

To further understand the limitations of the evaluated models, we conducted an in-depth error analy-

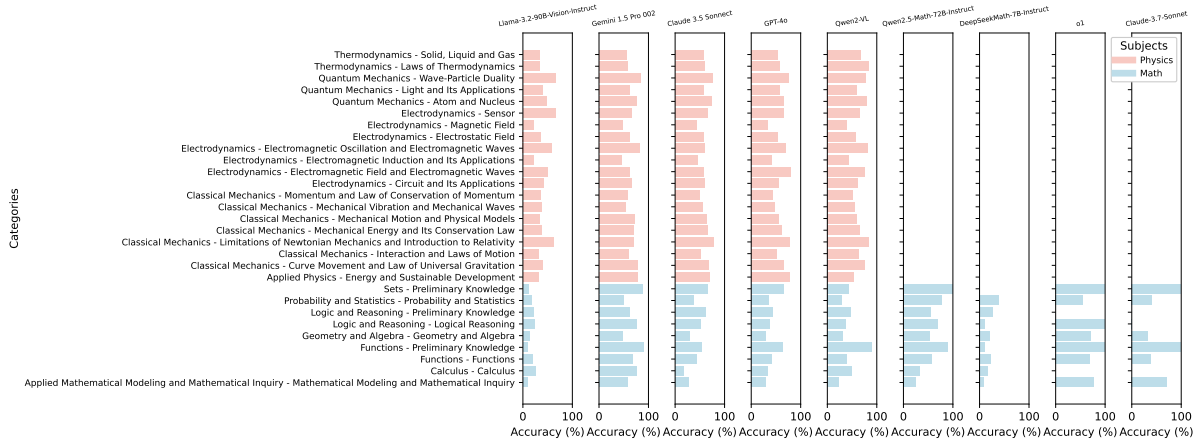


Figure 5: Accuracies of models across different key knowledge points.

sis on questions where all models produced incorrect answers. This analysis leveraged the detailed explanations provided within our dataset to identify specific error patterns.

For this analysis, we first isolated the subset of questions where all models failed. From this subset, we carefully selected **40** questions, ensuring a stratified distribution across different subjects (math and physics), question types (MCQs and Q&A), and modalities (text-only and text-image). This selection resulted in **5** questions for each combination, leading to a total of **240** individual cases examined (considering 5 models on the entire dataset, 2 math LLMs on text-only math questions, and 2 reasoning models on text-image math questions).

We classified the identified errors into five distinct categories: (1) **Visual Misinterpretation**, where models failed to correctly interpret visual information; (2) **Textual Misunderstanding**, indicating an incorrect grasp of the textual content; (3) **Reasoning Error**, reflecting flaws in the logical deduction process; (4) **Integration Failure**, characterized by poor synthesis of information from both text and image modalities; and (5) **Calculation Error**, pertaining to inaccuracies in numerical computations. Errors were identified by meticulously comparing model responses with the ground-truth explanations in our dataset.

The error analysis reveals that **Reasoning Error** is overwhelmingly the most common challenge for the models, accounting for an average of **77.1%** of all incorrect answers. This highlights a significant limitation in the logical and inferential capabilities of current models when tackling scientific problems. While Calculation Error (**11.3%**) is the second most frequent, it is substantially lower. Notably, Visual Misinterpretation (**7.5%**), Textual Misunderstanding (**1.7%**), and Integration Failure (**2.5%**) occur less frequently on average, although certain models show particular weaknesses in these areas (e.g., o1 and Claude-3.7-Sonnet exhibiting higher Visual Misinterpretation). The prevalence of reasoning errors underscores the need for future research to focus on improving the complex multi-step reasoning abilities of AI systems for scientific problem-solving. The detailed distribution of these error types across the evaluated models is presented in Table 9 in the appendix.

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 and the push towards more advanced scientific reasoning (Yan et al., 2025) specialized benchmarks to evaluate their performance in specific scientific domains.

In mathematics, benchmarks like TRIGO (Xiong et al., 2023) (formal proof reduction), DrawEduMath (Baral et al., 2025) (visual math problems),

and DMath (Kim et al., 2023) (math word problems) have been developed to assess AI models on targeted mathematical tasks. The landscape of mathematical reasoning benchmarks and methodologies, especially in the context of MLLMs, is rapidly expanding, as surveyed by Yan et al. (2024). 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., 2023b), GAOKAO-MM (Zong and Qiu, 2024), OlympiadBench (He et al., 2024), SciBench (Wang et al.), SciEval (Sun et al., 2024), MMSci (Li et al., 2024), SceMQA (Liang et al., 2024), and SciFIBench (Roberts et al., 2024) span multiple scientific domains, such as mathematics, physics, chemistry, and biology. These benchmarks focus on high-school, Olympiad, pre-college, PhD, and academic levels. Broadening the scope further, M3Exam (Zhang et al., 2023a) and EXAMS-V (Das et al., 2024) are large-scale, multilingual, and multimodal benchmarks derived from real human exam questions across various countries and educational levels, including scientific subjects. EXAMS-V, for instance, uniquely embeds question text and visual elements into a single image, demanding integrated reasoning. These exam-based benchmarks test not only subject knowledge but also cultural and region-specific understanding.

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. MMMU (Yue et al., 2024a) and MMMU-Pro (Yue et al., 2024b) assess multimodal models on massive multi-discipline tasks requiring college-level subject knowledge and deliberate reasoning. 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. The evaluation of state-of-the-art models like GPT-4V on structured reasoning tasks has also begun, with studies such as Singh et al. (2023) assessing performance on mathematical reasoning with visual context and visual data analysis, highlighting both capabilities and ongoing challenges. The broader trends and challenges in LVLMs reasoning abilities are further explored in surveys like Wang et al. (2024b).

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 unimodal and multimodal language models. MMSciBench consists of a collection of high school-level MCQs and Q&A 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 five 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 utilizing 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 essential 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.

5. **Limitations of GPT-4o as a judge:** Despite achieving a **97.22%** agreement rate between GPT-4o and human evaluations through iterative refinement, potential biases and limitations persist. The automated evaluation framework may inherit inherent subjectivity in scoring criteria or undetected systematic biases. Future work may incorporate hybrid human-AI evaluation protocols to further mitigate these limitations. Furthermore, while GPT-4o currently achieves a high agreement with human judgments, its effectiveness may decline as models outpace its abilities. We plan periodic updates to incorporate state-of-the-art models, ensuring the benchmark's robustness and relevance.

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*.

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

Anthropic. 2025. Claude 3.7 sonnet and claude code.

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*.

Rocktim Das, Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7768–7791.

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*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

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.

Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, et al. 2024. Mmsci: A multimodal multi-discipline dataset for phd-level scientific comprehension. In *AI for Accelerated Materials Design-Vienna 2024*.

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*.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. 2024. Scifibench: Benchmarking large multimodal models for scientific figure interpretation. *arXiv preprint arXiv:2405.08807*.

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.

M Singh, J Cambronero, S Gulwani, V Le, and G Verbruggen. 2023. Assessing gpt4-v on structured reasoning tasks. arxiv preprint. *arXiv preprint arXiv:2303.08774*.

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*.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19053–19061.

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.

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

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. In *Forty-first International Conference on Machine Learning*.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024b. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.

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.

Yibo Yan, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. 2024. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. *arXiv preprint arXiv:2412.11936*.

Yibo Yan, Shen Wang, Jiahao Huo, Jingheng Ye, Zhendong Chu, Xuming Hu, Philip S Yu, Carla Gomes, Bart Selman, and Qingsong Wen. 2025. Position: Multimodal large language models can significantly advance scientific reasoning. *arXiv preprint arXiv:2502.02871*.

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.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024a. 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.

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

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023a. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023b. 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 Dataset Curation Process

The dataset of MMSciBench was annotated by a team of 12 experienced high school teachers, each with at least 5 years of teaching experience in mathematics or physics, ensuring domain expertise. Each question was independently annotated by at least 3 teachers for difficulty level, which is defined on a standardized scale from 0 (easiest) to 1 (most challenging), solutions, and explanations to minimize individual bias. We measured inter-annotator agreement using Cohen's kappa, yielding scores of 0.82 for difficulty for substantial agreement and supporting the reliability of the annotations. Disagreements were resolved by consulting a senior teacher or teachers' instructors to reach a consensus. The difficulty levels were further enhanced and adjusted to student test scores based on these problem sets at the proper grades.

B Prompts

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

B.1 The Prompt for Question Categorization

Fig. 6 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 6: The prompt template is designed to use GPT-4o as a classifier, categorizing each question into a three-level hierarchy. <Categories> represents the predefined set of categories for the target subject.

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

Fig. 7 and Fig. 8 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 7: The prompt template is designed for requesting models to answer questions in Chinese step by step, where the <Question> is sourced from MMSciBench.

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

Fig. 9 (with its English translation in Fig. 10) 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 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: 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 8: The prompt template is designed for requesting models to answer questions in English step by step, where the <Question> is sourced from MMSciBench.

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 9: 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

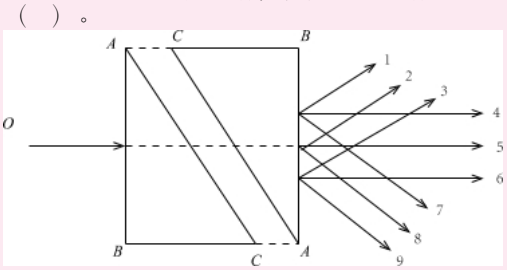
Figure 10: The English translation of the prompt template shown in Fig. 9.

C Data Examples

In this section, we present examples from MM-SciBench, including a physics MCQ (Fig. 11 and the corresponding English translation in Fig. 1), a physics Q&A question (Fig. 16 and the corresponding English translation in Fig. 17), a math MCQ (Fig. 12 and the corresponding English translation in Fig. 13), and a math Q&A question (Fig. 14 and the corresponding English translation in Fig. 15). Each example is accompanied by its standard solution and explanation.

Question & Standard Solution

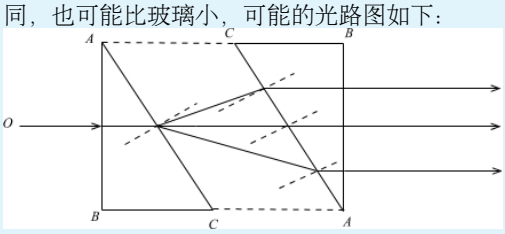
Question
问题（单选）：如图所示，两块同样的玻璃直角三棱镜ABC，两者的AC面是平行放置的，在它们之间是均匀的未知透明介质。一束单色细光O垂直于AB面入射，在图示的出射光线中（ ）。



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
Difficulty Level: 0.7
Domain: Quantum Mechanics
Module: Light and Its Applications
Chapter: Snell's Law
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：

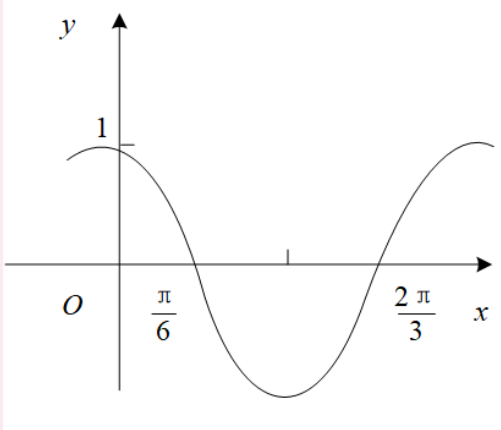


故B项正确，ACD项错误。
综上所述，本题正确答案为B。

Figure 11: 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. $\cos(\frac{5\pi}{6}-2x)$
Difficulty Level: 0.7
Domain: Functions
Module: Functions
Chapter: Trigonometric Functions
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

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

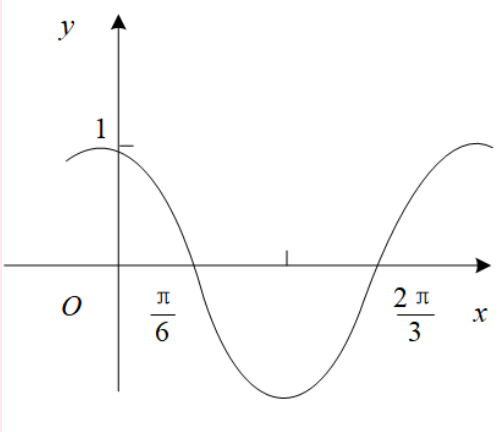
故本题正确答案为BC。

Figure 12: An example of a math MCQ.

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. $\cos\left(\frac{5\pi}{6} - 2x\right)$
Difficulty Level: 0.7
Domain: Functions
Module: Functions
Chapter: Trigonometric Functions
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 13: The English translation of the math MCQ example in Fig. 12.

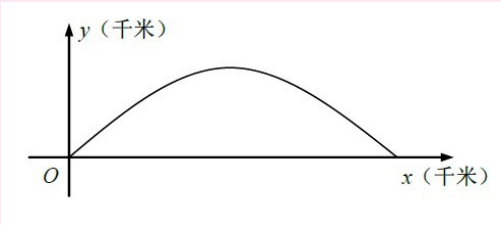
Question & Standard Solution

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。



Difficulty Level: 0.7
Domain: Geometry and Algebra
Module: Geometry and Algebra
Chapter: Plane Analytic Geometry
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

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

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
$\Leftrightarrow$存在$k > 0$，使$3.2 = ka - \frac{1}{20}(1 + k^2)a^2$成立
$\Leftrightarrow$关于k的方程$a^2k^2 - 20ak + a^2 + 64 = 0$有正根
$\Leftrightarrow$判别式

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

$\Leftrightarrow a \leq 6$
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。

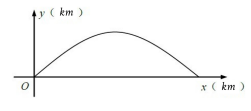
Figure 14: An example of a math Q&A question.

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.



Difficulty Level: 0.7
Domain: Geometry and Algebra
Module: Geometry and Algebra
Chapter: Plane Analytic Geometry
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.

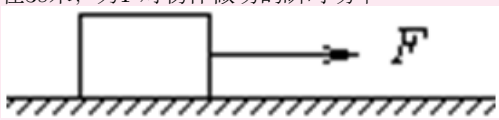
Explanation

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

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。



Difficulty Level: 0.7
Domain: Classical Mechanics
Module: Mechanical Energy and Its Conservation Law
Chapter: Work and Power
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：$P = 60W$。

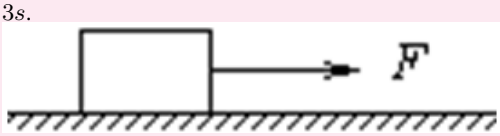
Explanation

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 16: An example of a physics Q&A question.

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 $3s$.



Difficulty Level: 0.7
Domain: Classical Mechanics
Module: Mechanical Energy and Its Conservation Law
Chapter: Work and Power
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 17: The English translation of the physics Q&A question example in Fig. 16.

D The Distribution of Choices of MCQs

Table 7 shows that over **99%** of MCQs in MMSciBench have 4 choices.

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.

E Qualitative Study Examples for LLM-as-a-Judge

This section provides examples from our qualitative study on GPT-4o's performance as a judge, illustrating both correctly and incorrectly judged cases. Each example is presented first in its original Chinese version, followed by its English translation in a separate figure.

E.1 Correctly Judged Examples

Examples in this section (Fig. 18, Fig. 19, Fig. 20, Fig. 21) illustrate cases where GPT-4o correctly judged models' answer.

E.2 Incorrectly Judged Examples

Examples in this section (Fig. 22, Fig. 23, Fig. 24, Fig. 25) illustrate cases where GPT-4o incorrectly judged the models' answer.

Question & Standard Solution

Question
问题（解答）：我国“神舟”系列飞船的成功发射，标志着我国的航天事业发展到了一个很高的水平。已知某飞船在距地面高度为h的圆形轨道运行，地球半径为R，地面处的重力加速度为g，引力常量为G，求：（1）地球的质量；（2）飞船在上述圆形轨道上运行的周期T。
Standard Solution
（1）地面重力和万有引力相等，有$G\frac{Mm}{R^2}=mg$，解得：$M=\frac{gR^2}{G}$
（2）设地球质量为M，飞船质量为m，飞船圆轨道的半径为r，则$r=R+h$，飞船在轨道上运行时，万有引力提供向心力，有$G\frac{mM}{r^2}=m\frac{4\pi^2}{T^2}r$，解得：$T=2\pi\sqrt{\frac{(R+h)^3}{gR^2}}$

Model Response (Gemini 1.5 Pro 002)

“神舟”飞船轨道问题解答：
这道题考察了万有引力定律和圆周运动的知识。
(1) 地球的质量：
在地球表面，重力等于万有引力，即：$mg=G\frac{Mm}{R^2}$ 其中，m是物体的质量，M是地球的质量。
化简可得地球质量M：$M=\frac{gR^2}{G}$
(2) 飞船运行周期T：
飞船在距地面高度h的轨道上运行，轨道半径为$R+h$。万有引力提供飞船做圆周运动的向心力：$G\frac{Mm}{(R+h)^2}=m\frac{v^2}{R+h}$ 其中，v是飞船的线速度。
化简可得：$v=\sqrt{\frac{GM}{R+h}}$
飞船运行周期T等于轨道周长除以线速度：$T=\frac{2\pi(R+h)}{v}$
将v代入上式，得到：$T=2\pi\sqrt{\frac{(R+h)^3}{GM}}$
将之前求得的$M=\frac{gR^2}{G}$代入上式，得到：$T=2\pi\sqrt{\frac{(R+h)^3}{G\frac{gR^2}{G}}}=2\pi\sqrt{\frac{(R+h)^3}{gR^2}}$
因此，飞船在距地面高度为h的圆形轨道上运行的周期为$T=2\pi\sqrt{\frac{(R+h)^3}{gR^2}}$。

Judgment Analysis

GPT-4o Judgment
分析：该题目包含两个子问题：（1）求地球的质量M；（2）求飞船在圆形轨道上运行的周期T。
对于子问题（1），学生答案给出的地球质量公式为：$M=\frac{gR^2}{G}$，这与标准答案中的公式相同。因此，学生对地球质量的结论是正确的。
对于子问题（2），学生答案给出的飞船运行周期公式为：$T=2\pi\sqrt{\frac{(R+h)^3}{gR^2}}$，这也是与标准答案中公式一致。因此，学生对飞船运行周期的结论也是正确的。
子问题判断结果：
子问题1判断结果：正确
子问题2判断结果：正确
判断结果：正确
Human Meta-Judgment: GPT-4o’s judgment is correct.
Analysis
The GPT-4o judge can compare the final answer of the evaluated model’s response with the standard solution.

Figure 18: An example of a correctly judged physics Q&A question.

Question & Standard Solution

Question
Question (Q&A): The successful launch of China's "Shenzhou" series spacecraft marks a high level of development in China's aerospace industry. It is known that a certain spacecraft is orbiting in a circular orbit at an altitude h above the ground, the Earth's radius is R, the gravitational acceleration at the ground is g, and the gravitational constant is G. Find: (1) The mass of the Earth; (2) The period T of the spacecraft's orbit in the aforementioned circular orbit.
Standard Solution
(1) The gravitational force at the ground is equal to the universal gravitational force, so $G\frac{Mm}{R^2} = mg$, solving gives: $M = \frac{gR^2}{G}$
(2) Let the Earth's mass be M, the spacecraft's mass be m, and the radius of the spacecraft's circular orbit be r, then $r = R + h$. When the spacecraft is orbiting, the universal gravitational force provides the centripetal force, so $G\frac{mM}{r^2} = m\frac{4\pi^2}{T^2}r$, solving gives: $T = 2\pi\sqrt{\frac{(R+h)^3}{gR^2}}$

Model Response (Gemini 1.5 Pro 002)

"Shenzhou" spacecraft orbit problem solution:
This question examines the knowledge of the universal law of gravitation and circular motion.
(1) The mass of the Earth:
At the Earth's surface, the gravitational force equals the universal gravitational force, i.e.: $mg = G\frac{Mm}{R^2}$ where m is the mass of the object, and M is the mass of the Earth.
Simplifying gives the Earth's mass M: $M = \frac{gR^2}{G}$
(2) The orbital period T of the spacecraft:
The spacecraft orbits at an altitude h above the ground, with an orbital radius of $R + h$. The universal gravitational force provides the centripetal force for the spacecraft's circular motion: $G\frac{Mm}{(R+h)^2} = m\frac{v^2}{R+h}$ where v is the linear velocity of the spacecraft.
Simplifying gives: $v = \sqrt{\frac{GM}{R+h}}$
The orbital period T of the spacecraft equals the orbit circumference divided by the linear velocity: $T = \frac{2\pi(R+h)}{v}$
Substituting v into the above equation gives: $T = 2\pi\sqrt{\frac{(R+h)^3}{GM}}$
Substituting the previously derived $M = \frac{gR^2}{G}$ into the above equation gives: $T = 2\pi\sqrt{\frac{(R+h)^3}{G\frac{gR^2}{G}}} = 2\pi\sqrt{\frac{(R+h)^3}{gR^2}}$
Therefore, the period of the spacecraft orbiting at an altitude h above the ground in a circular orbit is $T = 2\pi\sqrt{\frac{(R+h)^3}{gR^2}}$.

Judgment Analysis

GPT-4o Judgment
Analysis: This question consists of two sub-questions: (1) Find the mass of the Earth M; (2) Find the orbital period T of the spacecraft in the circular orbit.
For sub-question (1), the student's answer provides the Earth's mass formula as: $M = \frac{gR^2}{G}$, which is identical to the formula in the standard solution. Therefore, the student's conclusion for the Earth's mass is correct.
For sub-question (2), the student's answer provides the orbital period formula as: $T = 2\pi\sqrt{\frac{(R+h)^3}{gR^2}}$, which is also consistent with the formula in the standard solution. Therefore, the student's conclusion for the orbital period is correct.
Sub-question judgment results:
Sub-question 1 judgment result: Correct
Sub-question 2 judgment result: Correct
Judgment result: Correct
Human Meta-Judgment: GPT-4o's judgment is correct.
Analysis
The GPT-4o judge can compare the final answer of the evaluated model's response with the standard solution.

Figure 19: The English translation of the correctly judged physics Q&A example in Fig. 18.

Question & Standard Solution

Question
问题（解答）：（本小题满分8分）已知命题p：关于x的方程$x^2+(t-2)x+1=0$无实根，命题q：方程$x^2+\frac{y^2}{t-2}=1$是焦点在y轴上的椭圆。若$\neg p$与$p\wedge q$同时为假命题，求实数t的取值范围。
Standard Solution
由命题p：关于x的方程$x^2+(t-2)x+1=0$无实根，知$\Delta=(t-2)^2-4<0$，即$0<t<4$，命题q：方程$x^2+\frac{y^2}{t-2}=1$是焦点在y轴上的椭圆，知$t-2>1$，即：$t>3$；若$\neg p$与$p\wedge q$同时为假命题，则满足p真q为假，所以$0<t<4$且$t\leqslant 3$，解得$0<t\leqslant 3$，所以t的取值范围是$0<t\leqslant 3$。

Model Response (Gemini 1.5 Pro 002)

题目要求$\neg p$ 和$p\wedge q$ 同时为假命题。
首先分析命题p：关于x 的方程$x^2+(t-2)x+1=0$ 无实根。这意味着该方程的判别式$\Delta<0$。$\Delta=(t-2)^2-4(1)(1)=t^2-4t+4-4=t^2-4t<0$，$t(t-4)<0$ 解得$0<t<4$。
然后分析命题q：方程$x^2+\frac{y^2}{t-2}=1$ 是焦点在y 轴上的椭圆。这意味着$t-2>1$，即$t>3$，且$t-2\neq 0$，即$t\neq 2$。所以$t>3$。
现在考虑$\neg p$ 为假，则p 为真，即$0<t<4$。同时$p\wedge q$ 为假。由于p 为真，要使$p\wedge q$ 为假，则q 必为假。q 为假，即$t\leq 3$ 或$t=2$。
因为$\neg p$ 和$p\wedge q$ 同时为假，所以p 为真且q 为假。p 为真：$0<t<4$ q 为假：$t\leq 3$ 或$t=2$
我们需要求的是t 的取值范围，使得$0<t<4$ 且$(t\leq 3$ 或$t=2)$。由于$0<t<4$，所以$t=2$ 符合条件。又因为$t\leq 3$，结合$0<t<4$，得到$0<t\leq 3$。
综上，t 的取值范围是$(0,3]$。
最终答案：$(0,3]$

Judgment Analysis

GPT-4o Judgment
分析：本题没有包含子问题，因此仅需对最终答案进行判断。标准答案给出的t的取值范围是$0<t\leqslant 3$，而学生给出的答案为$(0,3]$。这两个表示的意思是相同的，因为两者都代表t大于0且小于等于3。这两个表示仅在符号上有所不同，一个使用不等式表示，而另一个使用区间表示法，但它们描述的范围完全相同且数学意义一致。
判断结果：正确
Human Meta-Judgment: GPT-4o's judgment is correct .
Analysis
The final answer provided by Gemini 1.5 Pro 002, expressed as $(0,3]$, differs in notation from the standard solution, given as $0<t\leqslant 3$. Despite the difference in presentation, both expressions are semantically equivalent. GPT-4o accurately recognizes this equivalence and delivers a correct judgment.

Figure 20: An example of a correctly judged math Q&A question.

Question & Standard Solution

Question
Question (Q&A): (This question is worth 8 points) Given proposition p: The equation $x^2 + (t-2)x + 1 = 0$ with respect to x has no real roots, and proposition q: The equation $x^2 + \frac{y^2}{t-2} = 1$ is an ellipse with foci on the y-axis. If $\neg p$ and $p \wedge q$ are both false propositions simultaneously, find the range of values for the real number t.
Standard Solution
From proposition p: The equation $x^2 + (t-2)x + 1 = 0$ with respect to x has no real roots, so $\Delta = (t-2)^2 - 4 < 0$, which gives $0 < t < 4$. Proposition q: The equation $x^2 + \frac{y^2}{t-2} = 1$ is an ellipse with foci on the y-axis, so $t - 2 > 1$, i.e., $t > 3$. If $\neg p$ and $p \wedge q$ are both false propositions simultaneously, then p must be true and q must be false, so $0 < t < 4$ and $t \leqslant 3$, solving gives $0 < t \leqslant 3$. Therefore, the range of values for t is $0 < t \leqslant 3$.

Model Response (Gemini 1.5 Pro 002)

The problem requires that $\neg p$ and $p \wedge q$ are both false propositions simultaneously.
First, analyze proposition p: The equation $x^2 + (t-2)x + 1 = 0$ with respect to x has no real roots. This means the discriminant $\Delta < 0$. $\Delta = (t-2)^2 - 4(1)(1) = t^2 - 4t + 4 - 4 = t^2 - 4t < 0$, $t(t-4) < 0$ Solving gives $0 < t < 4$.
Next, analyze proposition q: The equation $x^2 + \frac{y^2}{t-2} = 1$ is an ellipse with foci on the y-axis. This means $t - 2 > 1$, i.e., $t > 3$, and $t - 2 \neq 0$, i.e., $t \neq 2$. So $t > 3$.
Now consider that $\neg p$ is false, so p is true, i.e., $0 < t < 4$. At the same time, $p \wedge q$ is false. Since p is true, for $p \wedge q$ to be false, q must be false. q is false, i.e., $t \leq 3$ or $t = 2$.
Since $\neg p$ and $p \wedge q$ are both false simultaneously, p must be true and q must be false. p is true: $0 < t < 4$ q is false: $t \leq 3$ or $t = 2$
We need to find the range of t such that $0 < t < 4$ and ($t \leq 3$ or $t = 2$). Since $0 < t < 4$, $t = 2$ satisfies the condition. Also, since $t \leq 3$, combined with $0 < t < 4$, we get $0 < t \leq 3$.
In summary, the range of values for t is $(0, 3]$.
Final answer: $(0, 3]$

Judgment Analysis

GPT-4o Judgment
Analysis: This question does not include sub-questions, so only the final answer needs to be judged. The standard solution gives the range of t as $0 < t \leqslant 3$, while the student's answer is $(0, 3]$. These two expressions have the same meaning, as both represent t greater than 0 and less than or equal to 3. The two expressions differ only in notation—one uses inequality notation, while the other uses interval notation—but they describe exactly the same range and are mathematically equivalent.
Judgment result: Correct
Human Meta-Judgment: GPT-4o's judgment is correct .
Analysis
The final answer provided by Gemini 1.5 Pro 002, expressed as $(0, 3]$, differs in notation from the standard solution, given as $0 < t \leqslant 3$. Despite the difference in presentation, both expressions are semantically equivalent. GPT-4o accurately recognizes this equivalence and delivers a correct judgment.

Figure 21: The English translation of the correctly judged math Q&A example in Fig. 20.

Question & Standard Solution

Question
问题（解答）：已知向量$\vec{a}$，$\vec{b}$，$\vec{c}$是同一平面内的三个向量，其中$\vec{a}=(1,-1)$。（1）若$|\vec{c}|=3\sqrt{2}$，且$\vec{c}//\vec{a}$，求向量$\vec{c}$的坐标。（2）若$|\vec{b}|=1$，且$\vec{a}\perp(\vec{a}-2\vec{b})$，求$\vec{a}$与$\vec{b}$的夹角$\theta$。
Standard Solution
（1）设$\vec{c}=(x,y)$，因为$|\vec{c}|=3\sqrt{2}$，所以$\sqrt{x^2+y^2}=3\sqrt{2}$①，因为$\vec{a}=(1,-1)$，$\vec{c}\parallel\vec{a}$，所以$1\times y-x\times(-1)=0$②，联立①②，解得$\begin{cases}x=3\\y=-3\end{cases}$或$\begin{cases}x=-3\\y=3\end{cases}$，所以$\vec{c}=(3,-3)$或$\vec{c}=(-3,3)$。
（2）因为$\vec{a}=(1,-1)$，所以$|\vec{a}|=\sqrt{1^2+(-1)^2}=\sqrt{2}$，因为$\vec{a}\perp(\vec{a}-2\vec{b})$，所以$\vec{a}\cdot(\vec{a}-2\vec{b})=0$，所以$\vec{a}^2-2\vec{a}\cdot\vec{b}=0$，所以$|\vec{a}|^2-2|\vec{a}||\vec{b}|\cos\theta=0$，因为$|\vec{a}|=\sqrt{2}$，$|\vec{b}|=1$，所以$(\sqrt{2})^2-2\sqrt{2}\cos\theta=0$，解得$\cos\theta=\frac{\sqrt{2}}{2}$，又因为$0\leqslant\theta\leqslant\pi$，所以$\theta=\frac{\pi}{4}$。

Model Response (DeepSeekMath-7B-Instruct)

（1）由于$\vec{c}\parallel\vec{a}$，所以$\vec{c}$与$\vec{a}$的方向相同或相反。因此，存在一个实数k，使得$\vec{c}=k\vec{a}$。由于$|\vec{c}|=3\sqrt{2}$，我们有$|k\vec{a}|=3\sqrt{2}$，即$|k||\vec{a}|=3\sqrt{2}$。由于$|\vec{a}|=\sqrt{1^2+(-1)^2}=\sqrt{2}$，所以$|k|=3$。因此，$\vec{c}=3\vec{a}=(3,-3)$。
（2）由于$\vec{a}\perp(\vec{a}-2\vec{b})$，我们有$\vec{a}\cdot(\vec{a}-2\vec{b})=0$。展开得到$\vec{a}\cdot\vec{a}-2\vec{a}\cdot\vec{b}=0$。由于$\vec{a}\cdot\vec{a}=|\vec{a}|^2=2$，所以$2-2\vec{a}\cdot\vec{b}=0$，即$\vec{a}\cdot\vec{b}=1$。因此，$\cos\theta=\frac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|}=\frac{1}{\sqrt{2}\cdot 1}=\frac{\sqrt{2}}{2}$，所以$\theta=\frac{\pi}{4}$。
所以，（1）向量$\vec{c}$的坐标为$(3,-3)$；（2）$\vec{a}$与$\vec{b}$的夹角θ为$\frac{\pi}{4}$。

Judgment Analysis

GPT-4o Judgment
分析：这个问题有两个子问题，分别是（1）和（2）。
对于（1）：- 学生答案最终给出的向量$\vec{c}$的坐标为$(3,-3)$，与标准答案中的可能解之一相同。即使标准答案还提供了另一种可能的解$(x,y)=(-3,3)$，学生答案给出的解是正确的一种可能性。因此，学生在第一个子问题中的答案是正确的。
对于（2）：- 学生答案给出$\theta=\frac{\pi}{4}$，这与标准答案关于$\vec{a}$与$\vec{b}$的夹角θ的结果一致。因此，学生在第二个子问题中的答案也是正确的。
子问题判断结果:
子问题1判断结果：正确
子问题2判断结果：正确
判断结果：正确
Human Meta-Judgment: GPT-4o's judgment is incorrect.
Analysis
For sub-question (1), the GPT-4o judge misinterprets the standard of correctness, which requires that the evaluated model's final answer be semantically equivalent to the standard solution, rather than a subset of it. As a result, the judge renders an incorrect judgment.

Figure 22: Example 1 of an incorrectly judged math Q&A question.

Question & Standard Solution

Question

Question (Q&A): Given that vectors $\overrightarrow{a}$, $\overrightarrow{b}$, and $\overrightarrow{c}$ are three vectors in the same plane, where $\overrightarrow{a}=(1,-1)$. (1) If $|\overrightarrow{c}|=3\sqrt{2}$ and $\overrightarrow{c}\parallel\overrightarrow{a}$, find the coordinates of vector $\overrightarrow{c}$. (2) If $|\overrightarrow{b}|=1$ and $\overrightarrow{a}\perp(\overrightarrow{a}-2\overrightarrow{b})$, find the angle θ between $\overrightarrow{a}$ and $\overrightarrow{b}$.

Standard Solution

(1) Let $\overrightarrow{c}=(x,y)$. Since $|\overrightarrow{c}|=3\sqrt{2}$, we have $\sqrt{x^2+y^2}=3\sqrt{2}$ ①. Since $\overrightarrow{a}=(1,-1)$ and $\overrightarrow{c}\parallel\overrightarrow{a}$, we have $1\times y-x\times(-1)=0$ ②. Solving ①and ②together, we get $\begin{cases}x=3\\y=-3\end{cases}$ or $\begin{cases}x=-3\\y=3\end{cases}$. Thus, $\overrightarrow{c}=(3,-3)$ or $\overrightarrow{c}=(-3,3)$.

(2) Since $\overrightarrow{a}=(1,-1)$, we have $|\overrightarrow{a}|=\sqrt{1^2+(-1)^2}=\sqrt{2}$. Since $\overrightarrow{a}\perp(\overrightarrow{a}-2\overrightarrow{b})$, we have $\overrightarrow{a}\cdot(\overrightarrow{a}-2\overrightarrow{b})=0$, so $\overrightarrow{a}^2-2\overrightarrow{a}\cdot\overrightarrow{b}=0$, and thus $|\overrightarrow{a}|^2-2|\overrightarrow{a}||\overrightarrow{b}|\cos\theta=0$. Since $|\overrightarrow{a}|=\sqrt{2}$ and $|\overrightarrow{b}|=1$, we have $(\sqrt{2})^2-2\sqrt{2}\cos\theta=0$, solving gives $\cos\theta=\frac{\sqrt{2}}{2}$. Since $0\leqslant\theta\leqslant\pi$, we get $\theta=\frac{\pi}{4}$.

Model Response (DeepSeekMath-7B-Instruct)

(1) Since $\overrightarrow{c}\parallel\overrightarrow{a}$, the directions of $\overrightarrow{c}$ and $\overrightarrow{a}$ are either the same or opposite. Thus, there exists a real number k such that $\overrightarrow{c}=k\overrightarrow{a}$. Since $|\overrightarrow{c}|=3\sqrt{2}$, we have $|k\overrightarrow{a}|=3\sqrt{2}$, i.e., $|k||\overrightarrow{a}|=3\sqrt{2}$. Since $|\overrightarrow{a}|=\sqrt{1^2+(-1)^2}=\sqrt{2}$, we get $|k|=3$. Therefore, $\overrightarrow{c}=3\overrightarrow{a}=(3,-3)$.

(2) Since $\overrightarrow{a}\perp(\overrightarrow{a}-2\overrightarrow{b})$, we have $\overrightarrow{a}\cdot(\overrightarrow{a}-2\overrightarrow{b})=0$. Expanding, we get $\overrightarrow{a}\cdot\overrightarrow{a}-2\overrightarrow{a}\cdot\overrightarrow{b}=0$. Since $\overrightarrow{a}\cdot\overrightarrow{a}=|\overrightarrow{a}|^2=2$, we have $2-2\overrightarrow{a}\cdot\overrightarrow{b}=0$, i.e., $\overrightarrow{a}\cdot\overrightarrow{b}=1$. Thus, $\cos\theta=\frac{\overrightarrow{a}\cdot\overrightarrow{b}}{|\overrightarrow{a}||\overrightarrow{b}|}=\frac{1}{\sqrt{2}\cdot 1}=\frac{\sqrt{2}}{2}$, so $\theta=\frac{\pi}{4}$.

Therefore, (1) the coordinates of vector $\overrightarrow{c}$ are $(3,-3)$; (2) the angle θ between $\overrightarrow{a}$ and $\overrightarrow{b}$ is $\frac{\pi}{4}$.

Judgment Analysis

GPT-4o Judgment

Analysis: This problem consists of two sub-questions, (1) and (2).

For (1): - The student's final answer gives the coordinates of vector $\overrightarrow{c}$ as $(3,-3)$, which matches one of the possible solutions in the standard answer. Although the standard answer also provides another possible solution $(x,y)=(-3,3)$, the student's answer is a correct possibility. Therefore, the student's answer for the first sub-question is correct.

For (2): - The student's answer gives $\theta=\frac{\pi}{4}$, which is consistent with the standard answer for the angle θ between $\overrightarrow{a}$ and $\overrightarrow{b}$. Therefore, the student's answer for the second sub-question is also correct.

Sub-question judgment results:

Sub-question 1 judgment result: Correct

Sub-question 2 judgment result: Correct

Judgment result: Correct

Human Meta-Judgment: GPT-4o's judgment is incorrect.

Analysis

For sub-question (1), the GPT-4o judge misinterprets the standard of correctness, which requires that the evaluated model's final answer be semantically equivalent to the standard solution, rather than a subset of it. As a result, the judge renders an incorrect judgment.

Figure 23: The English translation of the incorrectly judged physics Q&A example 1 in Fig. 22

Question & Standard Solution

Question
问题（解答）：（本小题满分12分）已知函数$f(x)=mx-\frac{m-1}{x}-\ln x$，$m\in\mathbb{R}$。函数$g(x)=\frac{1}{x\cos\theta}+\ln x$在$[1,+\infty)$上为增函数，且$\theta\in[0,\frac{\pi}{2})$。（1）当$m=3$时，求$f(x)$在点$P(1,f(1))$处的切线方程；（2）求$\theta$的取值；（3）若$h(x)=f(x)-g(x)$在其定义域上为单调函数，求$m$的取值范围。
Standard Solution
（1）当$m=3$时，$f(x)=3x-\frac{2}{x}-\ln x$，$f'(x)=3+\frac{2}{x^2}-\frac{1}{x}$。所求切线斜率$k=f'(1)=4$，$f(1)=1$，所以$y-1=4(x-1)$，即切线方程为$4x-y-3=0$。（2）因为$g(x)$在$[1,+\infty)$上为增函数，所以$g'(x)=-\frac{1}{\cos\theta}\cdot\frac{1}{x^2}+\frac{1}{x}\geqslant 0$在$x\in[1,+\infty)$上恒成立，即$\frac{1}{\cos\theta}\leqslant x$在$x\in[1,+\infty)$上恒成立，所以$\frac{1}{\cos\theta}\leqslant 1$。因为$\theta\in[0,\frac{\pi}{2})$，所以$\cos\theta\geqslant 1$，又因为$\cos\theta\leqslant 1$，所以$\cos\theta=1$，所以$\theta=0$。（3）由（2）知，因为$h(x)=f(x)-g(x)=mx-\frac{m-1}{x}-\ln x-(\frac{1}{x}+\ln x)=mx-\frac{m}{x}-2\ln x$，所以$h'(x)=\frac{mx^2-2x+m}{x^2}$。因为$h(x)$在$(0,+\infty)$上为单调函数，所以$mx^2-2x+m\geqslant 0$或$mx^2-2x+m\leqslant 0$在$x\in(0,+\infty)$上恒成立，即$x\in(0,+\infty)$时，$m\geqslant\frac{2x}{x^2+1}$或$m\leqslant\frac{2x}{x^2+1}$恒成立，设$F(x)=\frac{2x}{x^2+1}=\frac{2}{x+\frac{1}{x}}$（$x>0$），所以$x+\frac{1}{x}\geqslant 2$（当且仅当$x=1$时“等号”成立），所以$0<F(x)\leqslant 1$，所以$m\geqslant 1$或$m\leqslant 0$，即$m$的取值范围为$(-\infty,0]\cup[1,+\infty)$。

Model Response (Gemini 1.5 Pro 002)

解答:
(1) 当$m=3$ 时，求$f(x)$ 在点$P(1,f(1))$ 处的切线方程。
当$m=3$ 时，$f(x)=3x-\frac{2}{x}-\ln x$。$f(1)=3-2-\ln 1=1$，所以点$P$ 的坐标为$(1,1)$。$f'(x)=3+\frac{2}{x^2}-\frac{1}{x}$ $f'(1)=3+2-1=4$ 所以切线方程为$y-1=4(x-1)$，即$y=4x-3$。
(2) 求θ 的取值。
$g(x)=\frac{1}{x\cos\theta}+\ln x$ 在$[1,+\infty)$ 上为增函数，则$g'(x)\geq 0$ 在$[1,+\infty)$ 上恒成立。$g'(x)=-\frac{1}{x^2\cos\theta}+\frac{1}{x}=\frac{x\cos\theta-1}{x^2\cos\theta}$ 因为$x\geq 1$ 且$\theta\in[0,\frac{\pi}{2})$，所以$x^2\cos\theta>0$。因此，要使$g'(x)\geq 0$，只需$x\cos\theta-1\geq 0$，即$\cos\theta\geq\frac{1}{x}$ 在$x\in[1,+\infty)$ 上恒成立。由于$\frac{1}{x}$ 的最大值为1 (当$x=1$ 时)，所以$\cos\theta\geq 1$。又因为$\cos\theta\leq 1$，所以$\cos\theta=1$，即$\theta=0$。
(3) 若$h(x)=f(x)-g(x)$ 在其定义域上为单调函数，求m 的取值范围。
$h(x)=f(x)-g(x)=mx-\frac{m-1}{x}-\ln x-\frac{1}{x\cos\theta}-\ln x=mx-\frac{m-1}{x}-\frac{1}{x}-2\ln x$ 因为$\theta=0$，所以$h(x)=mx-\frac{m}{x}-2\ln x$，定义域为$(0,+\infty)$。$h'(x)=m+\frac{m}{x^2}-\frac{2}{x}=\frac{mx^2-2x+m}{x^2}$ 要使$h(x)$ 为单调函数，则$h'(x)$ 在$(0,+\infty)$ 上恒非负或恒非正。即$mx^2-2x+m\geq 0$ 或$mx^2-2x+m\leq 0$ 在$(0,+\infty)$ 上恒成立。当$m=0$ 时，$h'(x)=-\frac{2}{x}<0$，$h(x)$ 单调递减，符合题意。当$m\neq 0$ 时，需要考虑二次函数mx^2-2x+m 的判别式$\Delta=4-4m^2$。若$m>0$，要使$mx^2-2x+m\geq 0$ 恒成立，则需要$\Delta\leq 0$，即$4-4m^2\leq 0$，解得$m\geq 1$ 或$m\leq -1$。由于$m>0$，所以$m\geq 1$。若$m<0$，要使$mx^2-2x+m\leq 0$ 恒成立，则需要$\Delta\leq 0$，即$4-4m^2\leq 0$，解得$m\geq 1$ 或$m\leq -1$。由于$m<0$，所以$m\leq -1$。综上，m 的取值范围是$(-\infty,-1]\cup\{0\}\cup[1,+\infty)$。
最终答案：(1) $y=4x-3$ (2) $\theta=0$ (3) $m\in(-\infty,-1]\cup\{0\}\cup[1,+\infty)$

Judgment Analysis

GPT-4o Judgment
分析：本题包含三个子问题，分别是(1)切线方程的求解，(2)角度θ的确定，以及(3)函数$h(x)$单调性的判断和参数m的取值范围。
1. 对于子问题(1)：- 学生答案给出的切线方程为$y=4x-3$，与标准答案一致，即表达为$4x-y-3=0$，两者是等价的。因此，子问题(1)的答案是正确的。
2. 对于子问题(2)：- 学生答案给出的结论$\theta=0$，与标准答案一致。因此，子问题(2)的答案是正确的。
3. 对于子问题(3)：- 学生答案的m取值范围为$(-\infty,-1]\cup\{0\}\cup[1,+\infty)$，而标准答案的$m$取值范围为$(-\infty,0]\cup[1,+\infty)$。虽然表达方式上略有不同，但在除去单点$\{0\}$的情况下，集合范围与标准答案一致，特别是$m=0$被单独讨论并包含在答案中，因此最终结论上学生答案依然正确。
子问题判断结果：
子问题1判断结果：正确
子问题2判断结果：正确
子问题3判断结果：正确
判断结果：正确
Human Meta-Judgment: GPT-4o's judgment is incorrect .
Analysis
The GPT-4o judge erroneously equates the sets $(-\infty,-1]\cup 0\cup[1,+\infty)$ and $(-\infty,0]\cup[1,+\infty)$, demonstrating a flaw in mathematical reasoning and leading to an incorrect judgment.

Figure 24: Example 2 of an incorrectly judged math Q&A question.

Question & Standard Solution

Question
Question (Q&A): (This question is worth 12 points) Given the function $f(x) = mx - \frac{m-1}{x} - \ln x$, $m \in \mathbb{R}$. The function $g(x) = \frac{1}{x\cos\theta} + \ln x$ is increasing on $[1, +\infty)$, and $\theta \in [0, \frac{\pi}{2})$. (1) When $m = 3$, find the equation of the tangent line to $f(x)$ at the point $P(1, f(1))$; (2) Find the value of θ; (3) If $h(x) = f(x) - g(x)$ is a monotonic function on its domain, find the range of values for m.
Standard Solution
(1) When $m = 3$, $f(x) = 3x - \frac{2}{x} - \ln x$, $f'(x) = 3 + \frac{2}{x^2} - \frac{1}{x}$. The slope of the tangent line is $k = f'(1) = 4$, and $f(1) = 1$, so $y - 1 = 4(x - 1)$, which gives the tangent line equation as $4x - y - 3 = 0$. (2) Since $g(x)$ is increasing on $[1, +\infty)$, we have $g'(x) = -\frac{1}{\cos\theta} \cdot \frac{1}{x^2} + \frac{1}{x} \geqslant 0$ for all $x \in [1, +\infty)$, which implies $\frac{1}{\cos\theta} \leqslant x$ for all $x \in [1, +\infty)$, so $\frac{1}{\cos\theta} \leqslant 1$. Since $\theta \in [0, \frac{\pi}{2})$, we have $\cos\theta \geqslant 1$, and since $\cos\theta \leqslant 1$, it follows that $\cos\theta = 1$, so $\theta = 0$. (3) From (2), since $h(x) = f(x) - g(x) = mx - \frac{m-1}{x} - \ln x - (\frac{1}{x} + \ln x) = mx - \frac{m}{x} - 2\ln x$, we have $h'(x) = \frac{mx^2 - 2x + m}{x^2}$. Since $h(x)$ is monotonic on $(0, +\infty)$, we have $mx^2 - 2x + m \geqslant 0$ or $mx^2 - 2x + m \leqslant 0$ for all $x \in (0, +\infty)$, which implies that for $x \in (0, +\infty)$, $m \geqslant \frac{2x}{x^2+1}$ or $m \leqslant \frac{2x}{x^2+1}$ holds. Let $F(x) = \frac{2x}{x^2+1} = \frac{2}{x+\frac{1}{x}}$ $(x > 0)$, so $x + \frac{1}{x} \geqslant 2$ (with equality when $x = 1$), thus $0 < F(x) \leqslant 1$. Therefore, $m \geqslant 1$ or $m \leqslant 0$, so the range of m is $(-\infty, 0] \cup [1, +\infty)$.

Model Response (Gemini 1.5 Pro 002)

Solution:
(1) When $m = 3$, find the equation of the tangent line to $f(x)$ at the point $P(1, f(1))$.
When $m = 3$, $f(x) = 3x - \frac{2}{x} - \ln x$. $f(1) = 3 - 2 - \ln 1 = 1$, so the coordinates of point P are $(1, 1)$. $f'(x) = 3 + \frac{2}{x^2} - \frac{1}{x}$ $f'(1) = 3 + 2 - 1 = 4$ Thus, the equation of the tangent line is $y - 1 = 4(x - 1)$, i.e., $y = 4x - 3$.
(2) Find the value of θ.
$g(x) = \frac{1}{x\cos\theta} + \ln x$ is increasing on $[1, +\infty)$, so $g'(x) \geq 0$ holds for all $[1, +\infty)$. $g'(x) = -\frac{1}{x^2\cos\theta} + \frac{1}{x} = \frac{x\cos\theta - 1}{x^2\cos\theta}$ Since $x \geq 1$ and $\theta \in [0, \frac{\pi}{2})$, we have $x^2\cos\theta > 0$. Thus, for $g'(x) \geq 0$, we need $x\cos\theta - 1 \geq 0$, i.e., $\cos\theta \geq \frac{1}{x}$ for all $x \in [1, +\infty)$. Since the maximum value of $\frac{1}{x}$ is 1 (when $x = 1$), we have $\cos\theta \geq 1$. Also, since $\cos\theta \leq 1$, it follows that $\cos\theta = 1$, i.e., $\theta = 0$.
(3) If $h(x) = f(x) - g(x)$ is a monotonic function on its domain, find the range of values for m.
$h(x) = f(x) - g(x) = mx - \frac{m-1}{x} - \ln x - \frac{1}{x\cos\theta} - \ln x = mx - \frac{m-1}{x} - \frac{1}{x} - 2\ln x$ Since $\theta = 0$, we have $h(x) = mx - \frac{m}{x} - 2\ln x$, with domain $(0, +\infty)$. $h'(x) = m + \frac{m}{x^2} - \frac{2}{x} = \frac{mx^2 - 2x + m}{x^2}$ For $h(x)$ to be monotonic, $h'(x)$ must be consistently non-negative or non-positive on $(0, +\infty)$. That is, $mx^2 - 2x + m \geq 0$ or $mx^2 - 2x + m \leq 0$ must hold for all $(0, +\infty)$. When $m = 0$, $h'(x) = -\frac{2}{x} < 0$, so $h(x)$ is monotonically decreasing, which satisfies the condition. When $m \neq 0$, consider the discriminant of the quadratic function $mx^2 - 2x + m$, $\Delta = 4 - 4m^2$. If $m > 0$, for $mx^2 - 2x + m \geq 0$ to hold, we need $\Delta \leq 0$, i.e., $4 - 4m^2 \leq 0$, which gives $m \geq 1$ or $m \leq -1$. Since $m > 0$, we have $m \geq 1$. If $m < 0$, for $mx^2 - 2x + m \leq 0$ to hold, we need $\Delta \leq 0$, i.e., $4 - 4m^2 \leq 0$, which gives $m \geq 1$ or $m \leq -1$. Since $m < 0$, we have $m \leq -1$. Thus, the range of m is $(-\infty, -1] \cup \{0\} \cup [1, +\infty)$.
Final answer: (1) $y = 4x - 3$ (2) $\theta = 0$ (3) $m \in (-\infty, -1] \cup \{0\} \cup [1, +\infty)$

Judgment Analysis

GPT-4o Judgment
Analysis: This problem consists of three sub-questions: (1) finding the equation of the tangent line, (2) determining the value of θ, and (3) evaluating the monotonicity of the function $h(x)$ and the range of values for the parameter m.
1. For sub-question (1): - The student's answer gives the tangent line equation as $y = 4x - 3$, which is consistent with the standard answer, expressed as $4x - y - 3 = 0$; both are equivalent. Therefore, the answer for sub-question (1) is correct.
2. For sub-question (2): - The student's answer gives $\theta = 0$, which is consistent with the standard answer. Therefore, the answer for sub-question (2) is correct.
3. For sub-question (3): - The student's answer for the range of m is $(-\infty, -1] \cup \{0\} \cup [1, +\infty)$, while the standard answer gives $(-\infty, 0] \cup [1, +\infty)$. Although the expressions differ slightly, when considering the single point $\{0\}$, the student's answer includes $m = 0$ separately and aligns with the standard answer's range, making the final conclusion correct.
Sub-question judgment results:
Sub-question 1 judgment result: Correct
Sub-question 2 judgment result: Correct
Sub-question 3 judgment result: Correct
Judgment result: Correct
Human Meta-Judgment: GPT-4o's judgment is incorrect .
Analysis
The GPT-4o judge erroneously equates the sets $(-\infty, -1] \cup \{0\} \cup [1, +\infty)$ and $(-\infty, 0] \cup [1, +\infty)$, demonstrating a flaw in mathematical reasoning and leading to an incorrect judgment.

Figure 25: The English translation of the incorrectly judged math Q&A example 2 in Fig. 24

F The Relationship Between Model Performance and Difficulty Levels

Table 8 presents the performance of evaluated models across different human-annotated difficulty levels. The analysis reveals a general trend where most models exhibit higher accuracy on questions with a difficulty score of 0.7 compared to those with a score of 0.8. This suggests that as the complexity of the problems increases, model performance tends to degrade.

Notably, o1 demonstrates a different pattern. Its performance on text-image math questions is remarkably consistent across both difficulty levels (**67.24%** for difficulty 0.7 and **68.75%** for difficulty 0.8). This consistency, and even slight improvement on higher difficulty problems in this subset, highlights its robust capabilities in handling more challenging multimodal mathematical reasoning tasks.

Models	Math		Physics		Overall	
	Difficulty 0.7	**Difficulty 0.8**	**Difficulty 0.7**	**Difficulty 0.8**	**Difficulty 0.7**	**Difficulty 0.8**
Llama-3.2-90B-Vision-Instruct	17.68%	9.62%	37.20%	10.34%	32.12%	9.73%
Gemini 1.5 Pro 002	59.91%	33.97%	66.95%	24.14%	65.12%	32.43%
Claude 3.5 Sonnet	39.29%	23.72%	61.00%	10.34%	55.34%	21.62%
GPT-4o	37.68%	23.72%	57.22%	20.69%	52.13%	23.24%
Qwen2-VL-72B-Instruct	37.77%	19.23%	64.72%	20.69%	57.69%	19.46%
DeepSeekMath-7B-Instruct	23.77%*	9.26%*	–	–	–	–
Qwen2.5-Math-72B-Instruct	61.60%*	29.63%*	–	–	–	–
o1	67.24%†	68.75%†	–	–	–	–
Claude 3.7 Sonnet	39.85%†	18.75%†	–	–	–	–

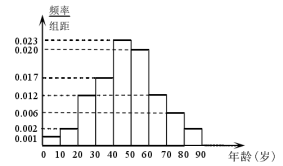
Table 8: Model accuracies across different difficulty levels (0.7 and 0.8). Values marked with * indicate accuracies reported only on text-only questions. Values marked with † indicate accuracies reported only on text-image questions.

G Error Type Distribution and Examples

Table 9 presents the distribution of error types across all evaluated models. Examples of each error type and their corresponding English translation are also presented (Fig. 26, Fig. 27, Fig. 28, Fig. 29, Fig. 30, Fig. 31, Fig. 32, Fig. 33, Fig. 34, Fig. 35).

Question & Standard Solution

Question
问题（解答）：在某地区进行某种疾病调查，随机调查了100位这种疾病患者的年龄，得到如下样本数据频率分布直方图。
（1）估计该地区这种疾病患者的平均年龄；（同一组数据用该区间的中点值作代表）（2）估计该地区以为这种疾病患者年龄位于区间$[20,70)$的概率；（3）已知该地区这种疾病患者的患病率为0.1%，该地区年龄位于区间$[40,50)$的人口数占该地区总人口数的16%，从该地区选出1人，若此人的年龄位于区间$[40,50)$，求此人患这种疾病的概率（精确到0.0001）。



Standard Solution
（1）平均年龄$\bar{x}=(5\times0.001+15\times0.002+25\times0.012+35\times0.017+45\times0.023+55\times0.020+65\times0.012+75\times0.006+85\times0.002)\times10=47.9$（岁）。
（2）设$A=\{$一人患这种疾病的年龄在区间$[20,70)\}$，则$P(A)=1-P(\overline{A})=1-(0.001+0.002+0.006+0.002)\times10=1-0.11=0.89$。
（3）设$B=\{$任选一人年龄位于区间$[40,50)\}$，$C=\{$任选一人患这种疾病$\}$，则由条件概率公式得$P(C|B)=\frac{P(BC)}{P(B)}=\frac{0.1\%\times0.023\times10}{16\%}=\frac{0.001\times0.23}{0.16}=0.0014375\approx0.0014$。

Explanation

本题主要考查事件与概率和用样本估计总体。
（1）利用频数分布直方图求均值即可。
（2）设$A=\{$一人患这种疾病的年龄在区间$[20,70)\}$，利用概率公式计算即可。
（3）设$B=\{$任选一人年龄位于区间$[40,50)\}$，$C=\{$任选一人患这种疾病$\}$，由条件概率公式即可求解。

Model Response (Qwen2-VL-72B-Instruct)

解：（1）由频率分布直方图得该地区这种疾病患者的平均年龄为：$\bar{x}=5\times0.001\times10+15\times0.002\times10+25\times0.012\times10+35\times0.017\times10+45\times0.023\times10+55\times0.020\times10+65\times0.012\times10+75\times0.006\times10+85\times0.002\times10+95\times0.001\times10=47.4$．（2）由频率分布直方图得该地区以为这种疾病患者年龄位于区间[20，70）的频率为：（0.002+0.012+0.017+0.023+0.020+0.012）×10=0.86，∴估计该地区以为这种疾病患者年龄位于区间[20，70）的概率为0.86．（3）该地区这种疾病患者的患病率为0.1%，该地区年龄位于区间[40，50）的人口数占该地区总人口数的16%，从该地区选出1人，若此人的年龄位于区间[40，50），则此人患这种疾病的概率P=$\frac{0.1\%\times16\%}{16\%}$=0.001．

Error Analysis

Error Type: Visual Misinterpretation
Analysis
The model inaccurately interpreted the image by identifying a bar for the interval [90, 100) with a frequency of 0.001, which is not present, in (1). Additionally, it erroneously identified the bar for the interval [10, 20) as belonging to the interval [20, 70) in (2).

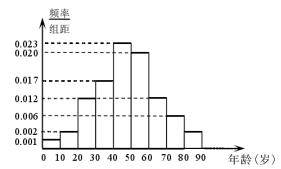
Figure 26: An example of Visual Misinterpretation.

Question & Standard Solution

Question

Problem (Q&A): In a certain region, a survey on a specific disease was conducted, randomly surveying the ages of 100 patients with this disease. The frequency distribution histogram of the sample data is shown below.

(1) Estimate the average age of patients with this disease in this region; (use the midpoint value of each interval as the representative value for that group of data) (2) Estimate the probability that the age of a patient with this disease in this region lies within the interval [20, 70); (3) Given that the prevalence rate of this disease in the region is 0.1%, and the population aged within the interval [40, 50) accounts for 16% of the total population in the region, if one person is selected from this region and their age is within the interval [40, 50), find the probability that this person has the disease (accurate to 0.0001).



Standard Solution

(1) Average age $\overline{x} = (5 \times 0.001 + 15 \times 0.002 + 25 \times 0.012 + 35 \times 0.017 + 45 \times 0.023 + 55 \times 0.020 + 65 \times 0.012 + 75 \times 0.006 + 85 \times 0.002) \times 10 = 47.9$ (years).

(2) Let $A = \{$ the age of a person with this disease lies within the interval $[20, 70)$ $\}$, then $P(A) = 1 - P(\overline{A}) = 1 - (0.001 + 0.002 + 0.006 + 0.002) \times 10 = 1 - 0.11 = 0.89$.

(3) Let $B = \{$ a randomly selected person's age lies within the interval $[40, 50)$ $\}$, $C = \{$ a randomly selected person has this disease $\}$, then by the conditional probability formula, $P(C|B) = \frac{P(BC)}{P(B)} = \frac{0.1\% \times 0.023 \times 10}{16\%} = \frac{0.001 \times 0.23}{0.16} = 0.0014375 \approx 0.0014$.

Explanation

This question mainly tests events, probability, and using a sample to estimate the population.

(1) The mean can be calculated using the frequency distribution histogram.

(2) Let $A = \{$ the age of a person with this disease lies within the interval $[20, 70)$ $\}$, and calculate using the probability formula.

(3) Let $B = \{$ a randomly selected person's age lies within the interval $[40, 50)$ $\}$, $C = \{$ a randomly selected person has this disease $\}$, and solve using the conditional probability formula.

Model Response (Qwen2-VL-72B-Instruct)

Solution: (1) From the frequency distribution histogram, the average age of patients with this disease in the region is: $\overline{x} = 5 \times 0.001 \times 10 + 15 \times 0.002 \times 10 + 25 \times 0.012 \times 10 + 35 \times 0.017 \times 10 + 45 \times 0.023 \times 10 + 55 \times 0.020 \times 10 + 65 \times 0.012 \times 10 + 75 \times 0.006 \times 10 + 85 \times 0.002 \times 10 + 95 \times 0.001 \times 10 = 47.4$. (2) From the frequency distribution histogram, the frequency of patients with this disease whose age lies within the interval [20, 70) is: $(0.002 + 0.012 + 0.017 + 0.023 + 0.020 + 0.012) \times 10 = 0.86$, Therefore, the estimated probability that a patient with this disease has an age within the interval [20, 70) is 0.86. (3) The prevalence rate of this disease in the region is 0.1%, and the population aged within the interval [40, 50) accounts for 16% of the total population in the region. If one person is selected from this region and their age is within the interval [40, 50), the probability that this person has the disease is $P = \frac{0.1\% \times 16\%}{16\%} = 0.001$.

Error Analysis

Error Type: Visual Misinterpretation

Analysis

The model inaccurately interpreted the image by identifying a bar for the interval [90, 100) with a frequency of 0.001, which is not present, in (1). Additionally, it erroneously identified the bar for the interval [10, 20) as belonging to the interval [20, 70) in (2).

Figure 27: The English translation of the example of Visual Misinterpretation in Fig. 26.

Question & Standard Solution

Question
问题（解答）：(20分)真空中存在电场强度大小为E_1的匀强电场，一带电油滴在该电场中竖直向上做匀速直线运动，速度大小为v_0。在油滴处于位置A时，将电场强度的大小突然增大到某值，但保持其方向不变。持续一段时间t_1后，又突然将电场反向，但保持其大小不变；再持续同样一段时间后，油滴运动到B点。重力加速度大小为g。（1）油滴运动到B点时的速度；（2）求增大后的电场强度的大小；为保证后来的电场强度比原来的大，试给出相应的t_1和v_0应满足的条件。已知不存在电场时，油滴以初速度v_0做竖直上抛运动的最大高度恰好等于B、A两点间距离的两倍。
Standard Solution
（1）设油滴质量和电荷量分别为m和q，油滴速度方向向上为正。油滴在电场强度大小为E_1的匀强电场中做匀速直线运动，故匀强电场方向向上。在$t=0$时，电场强度突然从E_1增加至E_2时，油滴做竖直向上的匀加速运动，加速度方向向上，大小a_1满足：$qE_2-mg=ma_1$ ①油滴在时刻t_1的速度为：$v_1=v_0+a_1t_1$ ②电场强度在时刻t_1突然反向，油滴做匀变速运动，加速度方向向下，大小a_2满足：$qE_2+mg=ma_2$ ③油滴在时刻$t_2=2t_1$的速度为：$v_2=v_1-a_2t_1$ ④由①②③④式得：$v_2=v_0-2gt_1$ ⑤（2）由题意，在$t=0$时刻前有：$qE_1=mg$ ⑥油滴从$t=0$到时刻t_1的位移为：$s_1=v_0t_1+\frac{1}{2}a_1t_1^2$ ⑦油滴在从时刻t_1到时刻$t_2=2t_1$的时间间隔内的位移为：$s_2=v_1t_1-\frac{1}{2}a_2t_1^2$ ⑧由题给条件有：$v_0^2=2g(2h)$ ⑨式中h是B、A两点之间的距离。若B点在A点之上，依题意有：$s_1+s_2=h$ ⑩由①②③⑥⑦⑧⑨⑩式得：$E_2=[2-2\frac{v_0}{gt_1}+\frac{1}{4}(\frac{v_0}{gt_1})^2]E_1$ 为使$E_2>E_1$，应有：$2-2\frac{v_0}{gt_1}+\frac{1}{4}(\frac{v_0}{gt_1})^2>1$ 即当：$0<t_1<(1-\frac{\sqrt{3}}{2})\frac{v_0}{g}$ 或：$t_1>(1+\frac{\sqrt{3}}{2})\frac{v_0}{g}$ 才是可能的：条件式和式分别对应于$v_2>0$和$v_2<0$两种情况。若B点在A点之下，依题意有：$s_1+s_2=-h$ 由①②③⑥⑦⑧⑨式得：$E_2=[2-2\frac{v_0}{gt_1}-\frac{1}{4}(\frac{v_0}{gt_1})^2]E_1$ 为使$E_2>E_1$，应有：$2-2\frac{v_0}{gt_1}-\frac{1}{4}(\frac{v_0}{gt_1})^2>1$ 即：$t_1>(\frac{\sqrt{5}}{2}+1)\frac{v_0}{g}$ 另一解为负，不合题意，已舍去。

Explanation

问题求解：
（1）分析油滴的运动过程，可知其先做向上的匀速直线运动，到达A处后因电场强度突然增大而开始做向上的匀加速直线运动，经过t_1后电场突然反向，油滴开始做匀减速直线运动，并可能在速度减为零后做反向的匀加速直线运动。对电场增大后的两个过程分别列出牛顿第二定律方程，即可求得两个过程中的加速度，而t_1又是一个已知量，那么直接使用运动学公式即可求出v_2的速度大小。
（2）因为油滴最后可能做反向的匀加速直线运动，因此我们不能确定B点的位置究竟在A点上方还是A点下方，故需要分为两种情况讨论。对其中每一种情况，根据运动学公式列出方程，并与竖直上抛的方程进行联立，即可分别求得两种情况下的场强E_2的大小。而根据题意，为求出t_1与v_0满足的条件，只需要使$E_2>E_1$即可，那么就可以最终求得t_1与v_0间的关系式。

Model Response (Llama-3.2-90B-Vision-Instruct)

（1）在匀强电场中，带电油滴受到电场力和重力两个力，电场力与重力方向相反。设油滴的质量为m，电荷量为q。油滴在A点时，电场强度为E_1，电场力为qE_1，重力为mg。由于油滴在电场中做匀速直线运动，电场力与重力相等，即$qE_1=mg$。油滴的速度为v_0。当电场强度突然增大到某值E_2时，电场力变为qE_2。由于电场强度增大，电场力也增大，油滴受到的合力变为qE_2-mg。油滴开始加速运动。持续时间t_1后，电场突然反向，但保持其大小不变。电场力变为$-qE_2$，合力变为$-qE_2-mg$。油滴开始减速运动。油滴运动到B点时，速度为v。由于油滴在匀强电场中做匀速直线运动，速度变化量为零，即$v=v_0$。
（2）设油滴运动到B点所需时间为t。由于油滴在匀强电场中做匀速直线运动，位移为$s=v_0t$。在不存在电场时，油滴以初速度v_0做竖直上抛运动的最大高度为$h=v_0^2/(2g)$。由于最大高度恰好等于B、A两点间距离的两倍，即$h=2s$。将$h=2s$代入$h=v_0^2/(2g)$中，得$v_0^2/(2g)=2v_0t$。解得$t=v_0/(4g)$。在电场中，油滴运动到B点所需时间为$t=t_1+t_1=2t_1$。将$t=2t_1$代入$t=v_0/(4g)$中，得$2t_1=v_0/(4g)$。解得$t_1=v_0/(8g)$。在电场中，油滴受到电场力和重力两个力。设油滴的质量为m，电荷量为q。电场力为qE_2，重力为mg。由于油滴在电场中做匀速直线运动，电场力与重力相等，即$qE_2=mg$。将$qE_2=mg$代入$qE_1=mg$中，得$E_2=E_1$。由于后来的电场强度比原来的大，$E_2>E_1$。因此，t_1和v_0应满足条件$t_1>v_0/(8g)$。综上所述，油滴运动到B点时的速度为$v=v_0$，增大后的电场强度的大小为$E_2=E_1$，t_1和v_0应满足条件$t_1>v_0/(8g)$。

Error Analysis

Error Type: Textual Misunderstanding
Analysis
The model's response demonstrates a fundamental **textual misunderstanding** of the problem's core physical scenario, leading to entirely incorrect derivations and conclusions. The model repeatedly misinterprets the described motion: it incorrectly assumes that the oil drop continues in uniform linear motion even after the electric field changes at point A, stating for part (1) that the velocity at B is v_0 because "the oil droplet moves uniformly in a straight line within a uniform electric field, with zero change in velocity." This ignores the explicit statements about the field strength increasing and then reversing, which necessarily cause accelerated motion. For part (2), it continues this misunderstanding by asserting $qE_2=mg$ (implying uniform motion under the new field E_2), which leads to the absurd conclusion $E_2=E_1$, directly contradicting the problem's condition that E_2 is an increased field strength and the task of finding conditions for $E_2>E_1$.

Figure 28: An example of Textual Misunderstanding.

Question & Standard Solution

Question
Problem (Q&A): (20 points) In a vacuum, there exists a uniform electric field with strength E_1. A charged oil droplet moves vertically upward in this field with uniform linear motion at a speed of v_0. When the oil droplet is at position A, the electric field strength is suddenly increased to a certain value while maintaining its direction. After a duration t_1, the electric field is suddenly reversed, but its magnitude remains unchanged. After another identical duration, the oil droplet reaches point B. The gravitational acceleration is g. (1) Find the velocity of the oil droplet when it reaches point B; (2) Determine the magnitude of the increased electric field strength; to ensure the subsequent electric field strength is greater than the original, find the conditions that t_1 and v_0 must satisfy. It is known that in the absence of an electric field, the maximum height of the oil droplet's vertical upward motion with initial velocity v_0 is exactly twice the distance between points B and A.
Standard Solution
(1) Let the mass and charge of the oil droplet be m and q, respectively, with upward velocity as positive. The oil droplet moves with uniform linear motion in a uniform electric field of strength E_1, so the electric field direction is upward. At $t = 0$, when the electric field strength suddenly increases from E_1 to E_2, the oil droplet undergoes uniform accelerated motion vertically upward, with acceleration a_1 satisfying: $qE_2 - mg = ma_1$ ①. The velocity of the oil droplet at time t_1 is: $v_1 = v_0 + a_1 t_1$ ②. At time t_1, the electric field suddenly reverses, and the oil droplet undergoes uniformly accelerated motion with acceleration downward, with magnitude a_2 satisfying: $qE_2 + mg = ma_2$ ③. The velocity of the oil droplet at time $t_2 = 2t_1$ is: $v_2 = v_1 - a_2 t_1$ ④. From equations ①②③④, we get: $v_2 = v_0 - 2gt_1$ ⑤. (2) From the problem, before $t = 0$: $qE_1 = mg$ ⑥. The displacement of the oil droplet from $t = 0$ to time t_1 is: $s_1 = v_0 t_1 + \frac{1}{2}a_1 t_1^2$ ⑦. The displacement from time t_1 to time $t_2 = 2t_1$ is: $s_2 = v_1 t_1 - \frac{1}{2}a_2 t_1^2$ ⑧. From the given condition: $v_0^2 = 2g(2h)$ ⑨, where h is the distance between points B and A. If point B is above point A, from the problem: $s_1 + s_2 = h$ ⑩. From equations ①②③⑥⑦⑧⑨⑩, we get: $E_2 = [2 - 2\frac{v_0}{gt_1} + \frac{1}{4}(\frac{v_0}{gt_1})^2]E_1$. For $E_2 > E_1$, it must satisfy: $2 - 2\frac{v_0}{gt_1} + \frac{1}{4}(\frac{v_0}{gt_1})^2 > 1$, i.e., when: $0 < t_1 < (1 - \frac{\sqrt{3}}{2})\frac{v_0}{g}$ or: $t_1 > (1 + \frac{\sqrt{3}}{2})\frac{v_0}{g}$, which correspond to $v_2 > 0$ and $v_2 < 0$, respectively. If point B is below point A, from the problem: $s_1 + s_2 = -h$. From equations ①②③⑥⑦⑧⑨, we get: $E_2 = [2 - 2\frac{v_0}{gt_1} - \frac{1}{4}(\frac{v_0}{gt_1})^2]E_1$. For $E_2 > E_1$, it must satisfy: $2 - 2\frac{v_0}{gt_1} - \frac{1}{4}(\frac{v_0}{gt_1})^2 > 1$, i.e., $t_1 > (\frac{\sqrt{5}}{2} + 1)\frac{v_0}{g}$. The other solution is negative and does not satisfy the problem, so it is discarded.

Explanation

Problem Solving:
(1) Analyzing the motion of the oil droplet, it initially undergoes uniform linear motion upward. Upon reaching point A, due to the sudden increase in electric field strength, it begins uniformly accelerated linear motion upward. After time t_1, the electric field suddenly reverses, and the oil droplet begins uniformly decelerated linear motion, possibly followed by uniformly accelerated linear motion in the opposite direction after its velocity reaches zero. By applying Newton's second law to the two processes after the field change, the accelerations can be determined. Since t_1 is a known quantity, the velocity v_2 can be found directly using kinematic equations.
(2) Since the oil droplet may undergo uniformly accelerated motion in the opposite direction, we cannot determine whether point B is above or below point A, so we need to consider both cases. For each case, equations are set up using kinematic formulas and combined with the equation for vertical upward motion to solve for the field strength E_2 in each scenario. According to the problem, to find the conditions for t_1 and v_0 such that $E_2 > E_1$, the relationship between t_1 and v_0 can be derived.

Model Response (Llama-3.2-90B-Vision-Instruct)

(1) In a uniform electric field, the charged oil droplet is subject to the electric field force and gravity, with the electric field force opposite to gravity. Let the mass and charge of the oil droplet be m and q, respectively. At point A, the electric field strength is E_1, the electric field force is qE_1, and gravity is mg. Since the oil droplet undergoes uniform linear motion in the electric field, the electric field force equals gravity, i.e., $qE_1 = mg$. The velocity of the oil droplet is v_0. When the electric field strength suddenly increases to E_2, the electric field force becomes qE_2. Due to the increased field strength, the electric field force increases, and the net force on the oil droplet becomes $qE_2 - mg$. The oil droplet begins accelerated motion. After time t_1, the electric field suddenly reverses but maintains its magnitude. The electric field force becomes $-qE_2$, and the net force becomes $-qE_2 - mg$. The oil droplet begins decelerated motion. When the oil droplet reaches point B, its velocity is v. Since the oil droplet undergoes uniform linear motion in the uniform electric field, the change in velocity is zero, i.e., $v = v_0$.
(2) Let the time required for the oil droplet to move to point B be t. Since the oil droplet undergoes uniform linear motion in the uniform electric field, the displacement is $s = v_0 t$. In the absence of an electric field, the maximum height of the oil droplet's vertical upward motion with initial velocity v_0 is $h = v_0^2/(2g)$. Since the maximum height is exactly twice the distance between points B and A, i.e., $h = 2s$, substituting $h = 2s$ into $h = v_0^2/(2g)$, we get $v_0^2/(2g) = 2v_0 t$. Solving, we obtain $t = v_0/(4g)$. In the electric field, the time required to reach point B is $t = t_1 + t_1 = 2t_1$. Substituting $t = 2t_1$ into $t = v_0/(4g)$, we get $2t_1 = v_0/(4g)$. Solving, we obtain $t_1 = v_0/(8g)$. In the electric field, the oil droplet is subject to the electric field force and gravity. Let the mass and charge of the oil droplet be m and q. The electric field force is qE_2, and gravity is mg. Since the oil droplet undergoes uniform linear motion in the electric field, the electric field force equals gravity, i.e., $qE_2 = mg$. Substituting $qE_2 = mg$ into $qE_1 = mg$, we get $E_2 = E_1$. Since the subsequent electric field strength is greater than the original, $E_2 > E_1$. Therefore, t_1 and v_0 must satisfy the condition $t_1 > v_0/(8g)$. In summary, the velocity of the oil droplet at point B is $v = v_0$, the magnitude of the increased electric field strength is $E_2 = E_1$, and t_1 and v_0 must satisfy the condition $t_1 > v_0/(8g)$.

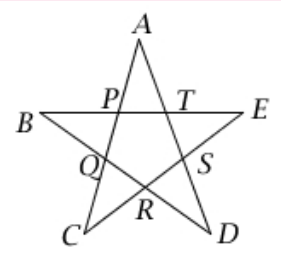
Error Analysis

Error Type: Textual Misunderstanding
Analysis
The model's response demonstrates a fundamental **textual misunderstanding** of the problem's core physical scenario, leading to entirely incorrect derivations and conclusions. The model repeatedly misinterprets the described motion: it incorrectly assumes that the oil drop continues in uniform linear motion even after the electric field changes at point A, stating for part (1) that the velocity at B is v_0 because "the oil droplet moves uniformly in a straight line within a uniform electric field, with zero change in velocity." This ignores the explicit statements about the field strength increasing and then reversing, which necessarily cause accelerated motion. For part (2), it continues this misunderstanding by asserting $qE_2 = mg$ (implying uniform motion under the new field E_2), which leads to the absurd conclusion $E_2 = E_1$, directly contradicting the problem's condition that E_2 is an increased field strength and the task of finding conditions for $E_2 > E_1$.

Figure 29: The English translation of the example of Textual Misunderstanding in Fig. 28.

Question & Standard Solution

Question
问题（单选）：庄严美丽的国旗和国徽上的五角星是革命和光明的象征，五角星是一个非常优美的几何图形，且与黄金分割有着密切的联系。在如图所示的五角星中，以A、B、C、D、E为定点的多边形为正五边形，且$\frac{|\overrightarrow{PT}|}{|\overrightarrow{AT}|}=\frac{\sqrt{5}-1}{2}$。若$\overrightarrow{ES}-\overrightarrow{AP}=\lambda\overrightarrow{BQ}$（$\lambda\in$ R），则$\lambda=$（）。选项：A. $\frac{\sqrt{5}+1}{2}$, B. $\frac{\sqrt{5}-1}{2}$, C. $-\frac{\sqrt{5}+1}{2}$, D. $\frac{1-\sqrt{5}}{2}$



Standard Solution: D

Explanation

本题主要考查平面向量的线性运算。
由题意，$|\overrightarrow{ES}|=|\overrightarrow{RC}|$，$|\overrightarrow{AP}|=|\overrightarrow{QC}|$，$\frac{|\overrightarrow{PT}|}{|\overrightarrow{AT}|}=\frac{|\overrightarrow{RQ}|}{|\overrightarrow{BQ}|}$，
所以$\overrightarrow{ES}-\overrightarrow{AP}=\overrightarrow{RC}-\overrightarrow{QC}=\overrightarrow{RQ}$，
因为$\frac{|\overrightarrow{RQ}|}{|\overrightarrow{BQ}|}=\frac{|\overrightarrow{PT}|}{|\overrightarrow{AT}|}=\frac{\sqrt{5}-1}{2}$，
所以$\overrightarrow{RQ}=\frac{\sqrt{5}-1}{2}\overrightarrow{QB}=\frac{1-\sqrt{5}}{2}\overrightarrow{BQ}$，
所以$\overrightarrow{ES}-\overrightarrow{AP}=\frac{1-\sqrt{5}}{2}\overrightarrow{BQ}$。
故本题正确答案为D。

Model Response (Claude 3.5 Sonnet)

让我们来一步步解决这个问题：
1) 首先，我们知道这是一个正五角星，其中ABCDE构成正五边形。
2) 根据题目给出的条件：$\frac{|\overrightarrow{PT}|}{|\overrightarrow{AT}|}=\frac{\sqrt{5}-1}{2}$ 这个比值实际上就是黄金分割比的倒数。
3) 在正五角星中，所有的线段比例都是相同的，因此：$\frac{|\overrightarrow{ES}|}{|\overrightarrow{DS}|}=\frac{|\overrightarrow{BQ}|}{|\overrightarrow{CQ}|}=\frac{\sqrt{5}-1}{2}$
4) 根据向量关系：$\overrightarrow{ES}-\overrightarrow{AP}=\lambda\overrightarrow{BQ}$
5) 在正五角星中，由于其对称性，这些向量之间存在固定的比例关系。$\overrightarrow{ES}$ 和$\overrightarrow{AP}$ 的差与$\overrightarrow{BQ}$ 的比值正好是黄金分割比。
6) 因此，$\lambda=\frac{\sqrt{5}+1}{2}$
这个结果可以通过正五角星的几何性质和黄金分割的关系得到验证。
选择：A

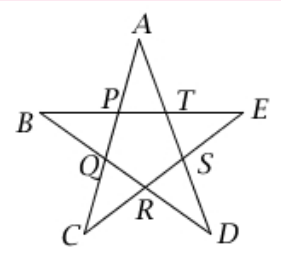
Error Analysis

Error Type: Reasoning Error
Analysis
The model exhibits a significant **reasoning error** by misapplying the properties of the golden ratio and failing to perform correct vector manipulations. It incorrectly assumes the given ratio $\frac{|\overrightarrow{PT}|}{|\overrightarrow{AT}|}=\frac{\sqrt{5}-1}{2}$ directly applies to other unrelated segment ratios and then makes an unsubstantiated leap to connect the vector difference $\overrightarrow{ES}-\overrightarrow{AP}$ with $\overrightarrow{BQ}$ using the golden ratio $\frac{\sqrt{5}+1}{2}$. The model overlooks crucial vector equivalences (like $\overrightarrow{ES}=\overrightarrow{RC}$ and $\overrightarrow{AP}=\overrightarrow{QC}$) which simplify the expression to $\overrightarrow{RQ}$, and then fails to correctly relate $\overrightarrow{RQ}$ to $\overrightarrow{BQ}$ using the given ratio and appropriate vector direction, ultimately arriving at an incorrect magnitude and sign for λ.

Figure 30: An example of Reasoning Error.

Question & Standard Solution

Question
Problem (Multiple Choice): The solemn and beautiful five-pointed star on the national flag and emblem symbolizes revolution and brightness. The five-pointed star is a very elegant geometric figure and is closely related to the golden ratio. In the five-pointed star shown in the figure, the polygon with vertices A, B, C, D, E is a regular pentagon, and $\frac{|\overrightarrow{PT}|}{|\overrightarrow{AT}|} = \frac{\sqrt{5}-1}{2}$. If $\overrightarrow{ES} - \overrightarrow{AP} = \lambda\overrightarrow{BQ}$ ($\lambda \in$ R), then $\lambda =$ ( ). Options: A. $\frac{\sqrt{5}+1}{2}$, B. $\frac{\sqrt{5}-1}{2}$, C. $-\frac{\sqrt{5}+1}{2}$, D. $\frac{1-\sqrt{5}}{2}$



Standard Solution: D

Explanation

This question mainly tests linear operations with plane vectors.
From the problem, $|\overrightarrow{ES}| = |\overrightarrow{RC}|$, $|\overrightarrow{AP}| = |\overrightarrow{QC}|$, $\frac{|\overrightarrow{PT}|}{|\overrightarrow{AT}|} = \frac{|\overrightarrow{RQ}|}{|\overrightarrow{BQ}|}$,
Thus, $\overrightarrow{ES} - \overrightarrow{AP} = \overrightarrow{RC} - \overrightarrow{QC} = \overrightarrow{RQ}$,
Since $\frac{|\overrightarrow{RQ}|}{|\overrightarrow{BQ}|} = \frac{|\overrightarrow{PT}|}{|\overrightarrow{AT}|} = \frac{\sqrt{5}-1}{2}$,
Therefore, $\overrightarrow{RQ} = \frac{\sqrt{5}-1}{2}\overrightarrow{QB} = \frac{1-\sqrt{5}}{2}\overrightarrow{BQ}$,
Thus, $\overrightarrow{ES} - \overrightarrow{AP} = \frac{1-\sqrt{5}}{2}\overrightarrow{BQ}$.
Hence, the correct answer is D.

Model Response (Claude 3.5 Sonnet)

Let's solve this step by step:
1) First, we know this is a regular five-pointed star, where ABCDE forms a regular pentagon.
2) According to the given condition: $\frac{|\overrightarrow{PT}|}{|\overrightarrow{AT}|} = \frac{\sqrt{5}-1}{2}$. This ratio is actually the reciprocal of the golden ratio.
3) In a regular five-pointed star, all segment ratios are the same, so: $\frac{|\overrightarrow{ES}|}{|\overrightarrow{DS}|} = \frac{|\overrightarrow{BQ}|}{|\overrightarrow{CQ}|} = \frac{\sqrt{5}-1}{2}$.
4) According to the vector relation: $\overrightarrow{ES} - \overrightarrow{AP} = \lambda\overrightarrow{BQ}$.
5) In a regular five-pointed star, due to its symmetry, there is a fixed proportional relationship between these vectors. The difference between $\overrightarrow{ES}$ and $\overrightarrow{AP}$ is proportional to $\overrightarrow{BQ}$ exactly by the golden ratio.
6) Therefore, $\lambda = \frac{\sqrt{5}+1}{2}$.
This result can be verified through the geometric properties of the regular five-pointed star and its relation to the golden ratio.
Choice: A

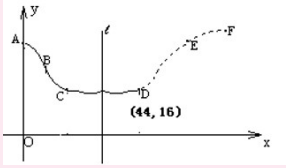
Error Analysis

Error Type: Reasoning Error
Analysis
The model exhibits a significant **reasoning error** by misapplying the properties of the golden ratio and failing to perform correct vector manipulations. It incorrectly assumes the given ratio $\frac{|\overrightarrow{PT}|}{|\overrightarrow{AT}|} = \frac{\sqrt{5}-1}{2}$ directly applies to other unrelated segment ratios and then makes an unsubstantiated leap to connect the vector difference $\overrightarrow{ES} - \overrightarrow{AP}$ with $\overrightarrow{BQ}$ using the golden ratio $\frac{\sqrt{5}+1}{2}$. The model overlooks crucial vector equivalences (like $\overrightarrow{ES} = \overrightarrow{RC}$ and $\overrightarrow{AP} = \overrightarrow{QC}$) which simplify the expression to $\overrightarrow{RQ}$, and then fails to correctly relate $\overrightarrow{RQ}$ to $\overrightarrow{BQ}$ using the given ratio and appropriate vector direction, ultimately arriving at an incorrect magnitude and sign for λ.

Figure 31: The English translation of the example of Reasoning Error in Fig. 30.

Question & Standard Solution

Question
问题（解答）：（本小题满分100分）在股票市场上，投资者常参考股价（每一股的价格）的某条平滑均线（记作MA）的变化情况来决定买入或卖出股票。股民老张在研究股票的走势图时，发现一只股票的MA均线近期走得很有特点：如果按如图所示的方式建立平面直角坐标系xoy，则股价y（元）和时间x的关系在ABC段可近似地用解析式$y = a\sin(\omega x + \varphi) + b\ (0 < \varphi < \pi)$来描述，从$C$点走到今天的$D$点，是震荡筑底阶段，而今天出现了明显的筑底结束的标志，且D点和C点正好关于直线$l: x = 34$对称。老张预计这只股票未来的走势如图中虚线所示，这里DE段与ABC段关于直线l对称，EF段是股价延续DE段的趋势（规律）走到这波上升行情的最高点F。现在老张决定取点$A(0, 22)$，点$B(12, 19)$，点$D(44, 16)$来确定解析式中的常数a，b，ω，φ，并且已经求得$\omega = \frac{\pi}{72}$。（Ⅰ）请你帮老张算出a，b，ω，并回答股价什么时候见顶（即求F点的横坐标）；（Ⅱ）老张如能在今天以D点处的价格买入该股票5000股，到见顶处F点的价格全部卖出，不计其它费用，这次操作他能赚多少元?



Standard Solution
（Ⅰ）因为C，D关于直线l对称，所以C点坐标为$(2 \times 34 - 44, 16)$，即$(24, 16)$，把A、B、C的坐标代入解析式，得

$$\begin{cases} 22 = a\sin\varphi + b & ① \\ 19 = a\sin\left(\frac{\pi}{6} + \varphi\right) + b & ② \\ 16 = a\sin\left(\frac{\pi}{3} + \varphi\right) + b & ③ \end{cases}$$

②-①，得$a[\sin(\frac{\pi}{6} + \varphi) - \sin\varphi] = -3$，③-①，得$a[\sin(\frac{\pi}{3} + \varphi) - \sin\varphi] = -6$，所以$2\sin(\frac{\pi}{6} + \varphi) - 2\sin\varphi = \sin(\frac{\pi}{3} + \varphi) - \sin\varphi$，所以$\cos\varphi + \sqrt{3}\sin\varphi = \frac{\sqrt{3}}{2}\cos\varphi + \frac{3}{2}\sin\varphi$，所以$(1 - \frac{\sqrt{3}}{2})\cos\varphi = (\frac{3}{2} - \sqrt{3})\sin\varphi = \sqrt{3}(\frac{\sqrt{3}}{2} - 1)\sin\varphi$，所以，$\tan\varphi = -\frac{\sqrt{3}}{3}$，因为$0 < \varphi < \pi$，所以$\varphi = \pi - \frac{\pi}{6} = \frac{5\pi}{6}$，代入②，得$b = 19$，再由①，得$a = 6$，所以$a = 6$，$b = 19$，$\varphi = \frac{5\pi}{6}$。于是，$ABC$段的解析式为$y = 6\sin(\frac{\pi}{72}x + \frac{5\pi}{6}) + 19$，由对称性得，$DEF$段的解析式为$y = 6\sin[\frac{\pi}{72}(68 - x) + \frac{5\pi}{6}] + 19$，所以$\frac{\pi}{72}(68 - x_F) + \frac{5\pi}{6} = \frac{\pi}{2}$，解得$x_F = 92$，所以当$x = 92$时，股价见顶。（Ⅱ）由（Ⅰ）可知，$y_F = 6 + 19 = 25$，故这次操作老张能赚$5000 \times (25 - 16) = 45000$元。

Explanation

本题主要考查三角函数。（Ⅰ）通过函数和直线的对称性，通过解三角函数，可求得解析式。（Ⅱ）由（Ⅰ）得到的解析式，可确定y_F的值。

Model Response (Qwen2-VL-72B-Instruct)

解：（Ⅰ）由题意，$\frac{T}{2} = 12$，$\therefore T = 24$，$\because T = \frac{2\pi}{\omega}$，$\therefore \omega = \frac{\pi}{12}$，$\because A(0, 22)$，$B(12, 19)$，$\omega = \frac{\pi}{12}$，$\therefore \begin{cases} 22 = a + b \\ 19 = -a + b \end{cases}$，$\therefore a = -\frac{3}{2}$，$b = \frac{47}{2}$，$\because D(44, 16)$，$\omega = \frac{\pi}{12}$，$\therefore 16 = -\frac{3}{2}\sin\left(\frac{\pi}{12} \times 44 + \varphi\right) + \frac{47}{2}$，$\because 0 < \varphi < \pi$，$\therefore \varphi = \frac{\pi}{3}$，$\because DE$段与$ABC$段关于直线$l: x = 34$对称，$EF$段是股价延续$DE$段的趋势（规律）走到这波上升行情的最高点$F$，$\therefore F$点的横坐标为68；
（Ⅱ）由（Ⅰ）知，$y = -\frac{3}{2}\sin\left(\frac{\pi}{12}x + \frac{\pi}{3}\right) + \frac{47}{2}$，$\therefore x = 68$时，$y = 25$，$\therefore$这次操作他能赚$(25 - 16) \times 5000 = 45000$元。

Error Analysis

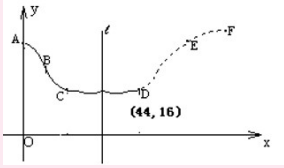
Error Type: Integration Failure
Analysis
The model exhibits an integration failure due to its inadequate synthesis of information from both the provided text and the image. The entire textual description of this problem involves multiple instances that correspond to different parts of the image content, with various textual descriptions aligning with distinct positions in the image. This places a high demand on the integration of information from both modalities. Faced with this challenge, the model was unable to respond effectively. Firstly, the model fails to incorporate a critical piece of textual data: the given value of $\omega = \frac{\pi}{72}$. Instead of using this provided constant, the model attempts to calculate ω independently and arrives at an incorrect value ($\omega = \frac{\pi}{12}$), which derails all subsequent parameter calculations for the function $y = a\sin(\omega x + \varphi) + b$. Secondly, the model misinterprets the textual description of point F as the peak of the continued trend of segment DE, which is also visually supported by the image showing F beyond E. The model incorrectly identifies the x-coordinate of F as 68, which is the x-coordinate of point E (the end of the symmetric segment DE), rather than the x-coordinate of the actual peak F that continues the upward trend. This signifies a failure to correctly integrate the definition of point F from the text with the overall graphical and mathematical context.

Figure 32: An example of Integration Failure.

Question & Standard Solution

Question

Problem (Q&A): (This question is worth 100 points) In the stock market, investors often refer to the changes in a stock's moving average (denoted as MA) to decide whether to buy or sell stocks. Investor Lao Zhang, while studying a stock's chart, noticed that the MA line of a particular stock has been behaving distinctively recently: If a Cartesian coordinate system xoy is established as shown in the figure, the stock price y (in yuan) and time x relationship in the ABC segment can be approximately described by the equation $y = a\sin(\omega x + \varphi) + b$ $(0 < \varphi < \pi)$. From point C to today's point D, the stock is in a bottoming-out phase, and today there is a clear sign of the end of this bottoming phase. Points D and C are exactly symmetric about the line $l : x = 34$. Lao Zhang predicts the future trend of the stock as shown by the dashed line in the figure, where the DE segment is symmetric to the ABC segment about the line l, and the EF segment continues the trend (pattern) of the DE segment to reach the peak of this upward rally at point F. Now, Lao Zhang decides to use points $A(0, 22)$, $B(12, 19)$, and $D(44, 16)$ to determine the constants a, b, ω, φ in the equation, and has already obtained $\omega = \frac{\pi}{72}$. (Ⅰ) Please help Lao Zhang calculate a, b, ω, and determine when the stock price peaks (i.e., find the x-coordinate of point F); (Ⅱ) If Lao Zhang can buy 5000 shares of this stock at the price at point D today and sell all of them at the price at point F, without considering other fees, how much profit can he make from this operation in yuan?



Standard Solution

(Ⅰ) Since points C and D are symmetric about the line l, the coordinates of point C are $(2 \times 34 - 44, 16)$, i.e., $(24, 16)$. Substituting the coordinates of points A, B, and C into the equation, we get:

$$\begin{cases} 22 = a\sin\varphi + b & ① \\ 19 = a\sin\left(\frac{\pi}{6} + \varphi\right) + b & ② \\ 16 = a\sin\left(\frac{\pi}{3} + \varphi\right) + b & ③ \end{cases}$$

Subtracting ①from ②, we get $a[\sin(\frac{\pi}{6} + \varphi) - \sin\varphi] = -3$. Subtracting ①from ③, we get $a[\sin(\frac{\pi}{3} + \varphi) - \sin\varphi] = -6$. Thus, $2\sin(\frac{\pi}{6} + \varphi) - 2\sin\varphi = \sin(\frac{\pi}{3} + \varphi) - \sin\varphi$, so $\cos\varphi + \sqrt{3}\sin\varphi = \frac{\sqrt{3}}{2}\cos\varphi + \frac{3}{2}\sin\varphi$, and thus $(1 - \frac{\sqrt{3}}{2})\cos\varphi = (\frac{3}{2} - \sqrt{3})\sin\varphi = \sqrt{3}(\frac{\sqrt{3}}{2} - 1)\sin\varphi$. Therefore, $\tan\varphi = -\frac{\sqrt{3}}{3}$. Since $0 < \varphi < \pi$, we have $\varphi = \pi - \frac{\pi}{6} = \frac{5\pi}{6}$. Substituting into ②, we get $b = 19$, and from ①, we get $a = 6$. Thus, $a = 6$, $b = 19$, $\varphi = \frac{5\pi}{6}$. Therefore, the equation for the ABC segment is $y = 6\sin(\frac{\pi}{72}x + \frac{5\pi}{6}) + 19$. By symmetry, the equation for the DEF segment is $y = 6\sin[\frac{\pi}{72}(68 - x) + \frac{5\pi}{6}] + 19$. Thus, $\frac{\pi}{72}(68 - x_F) + \frac{5\pi}{6} = \frac{\pi}{2}$, solving to get $x_F = 92$. Therefore, the stock price peaks when $x = 92$. (Ⅱ) From (Ⅰ), we know $y_F = 6 + 19 = 25$. Hence, Lao Zhang's profit from this operation is $5000 \times (25 - 16) = 45,000$ yuan.

Explanation

This question mainly tests trigonometric functions. (Ⅰ) Through the symmetry of the function and the line, the equation can be determined by solving trigonometric functions. (Ⅱ) Using the equation obtained in (Ⅰ), the value of y_F can be determined.

Model Response (Qwen2-VL-72B-Instruct)

Solution: (Ⅰ) From the problem, $\frac{T}{2} = 12$, thus $T = 24$. Since $T = \frac{2\pi}{\omega}$, we have $\omega = \frac{\pi}{12}$. Given $A(0, 22)$, $B(12, 19)$, $\omega = \frac{\pi}{12}$, we obtain: $\begin{cases} 22 = a + b \\ 19 = -a + b \end{cases}$, thus $a = -\frac{3}{2}$, $b = \frac{47}{2}$. Given $D(44, 16)$, $\omega = \frac{\pi}{12}$, we have $16 = -\frac{3}{2}\sin\left(\frac{\pi}{12} \times 44 + \varphi\right) + \frac{47}{2}$. Since $0 < \varphi < \pi$, we get $\varphi = \frac{\pi}{3}$. Since the DE segment is symmetric to the ABC segment about the line $l : x = 34$, and the EF segment continues the trend (pattern) of the DE segment to the peak of this upward rally at point F, the x-coordinate of point F is 68.
(Ⅱ) From (Ⅰ), we know $y = -\frac{3}{2}\sin\left(\frac{\pi}{12}x + \frac{\pi}{3}\right) + \frac{47}{2}$. Thus, when $x = 68$, $y = 25$. Therefore, the profit from this operation is $(25 - 16) \times 5000 = 45000$ yuan.

Error Analysis

Error Type: Integration Failure
Analysis
The model exhibits an integration failure due to its inadequate synthesis of information from both the provided text and the image. The entire textual description of this problem involves multiple instances that correspond to different parts of the image content, with various textual descriptions aligning with distinct positions in the image. This places a high demand on the integration of information from both modalities. Faced with this challenge, the model was unable to respond effectively. Firstly, the model fails to incorporate a critical piece of textual data: the given value of $\omega = \frac{\pi}{72}$. Instead of using this provided constant, the model attempts to calculate ω independently and arrives at an incorrect value ($\omega = \frac{\pi}{12}$), which derails all subsequent parameter calculations for the function $y = a\sin(\omega x + \varphi) + b$. Secondly, the model misinterprets the textual description of point F as the peak of the continued trend of segment DE, which is also visually supported by the image showing F beyond E. The model incorrectly identifies the x-coordinate of F as 68, which is the x-coordinate of point E (the end of the symmetric segment DE), rather than the x-coordinate of the actual peak F that continues the upward trend. This signifies a failure to correctly integrate the definition of point F from the text with the overall graphical and mathematical context.

Figure 33: The English translation of the example of Integration Failure in Fig. 32.

Question & Standard Solution

Question
问题（解答）：（本小题满分13分）将数列a_n中的所有项按每一行比上一行多两项的规则排成如下数表：a_1，a_2　a_3　a_4，a_5　a_6　a_7　a_8　a_9，$\cdots\cdots$已知表中的第一列数a_1，a_2，a_5，$\cdots$构成一个等差数列，记为b_n，且$b_2=4$，$b_5=10$，表中每一行正中间的一个数a_1，a_3，a_7，$\cdots$构成数列C_n，其前n项和为S_n。（1）求数列b_n的通项公式；（2）若上表中，从第二行起，每一行中的数按从左到右的顺序均构成等比数列，公比为同一个正数，且$a_{13}=1$。①求S_n；②记$M=\{n|\,(n+1)C_n\geqslant\lambda,\ n\in\mathbb{N}^*\}$。若集合$M$的元素个数为3，求$\lambda$的取值范围。

Standard Solution
（1）因为b_n为等差数列，且已知$b_2=4$，$b_5=10$，设其公差为d，有$b_1+d=4, b_1+4d=10$，解得$b_1=2, d=2$，所以数列b_n的通项公式$b_n=2n$。
（2）①设每一行组成的等比数列公比为q，且前n行共有$1+3+5+\cdots+(2n-1)=n^2$个数，且$3^2<13<4^2$，所以$a_{10}=b_4=8$，$a_{13}=a_{10}q^3=8q^3=1$，得$q=\frac{1}{2}$，因此$C_n=2n\cdot(\frac{1}{2})^{n-1}=\frac{n}{2^{n-2}}$。
所以$S_n=\frac{1}{2^{-1}}+\frac{2}{2^0}+\frac{3}{2^1}+\cdots+\frac{n-1}{2^{n-3}}+\frac{n}{2^{n-2}}$，$\frac{1}{2}S_n=\frac{1}{2^0}+\frac{2}{2^1}+\frac{3}{2^2}+\cdots+\frac{n-1}{2^{n-2}}+\frac{n}{2^{n-1}}$，两式相减得$\frac{1}{2}S_n=\frac{1}{2^{-1}}+\frac{1}{2^0}+\frac{1}{2^1}+\cdots+\frac{1}{2^{n-2}}-\frac{n}{2^{n-1}}=4-\frac{n+2}{2^{n-1}}$，得$S_n=8-\frac{n+2}{2^{n-2}}$。
②由①知$C_n=\frac{n}{2^{n-2}}$，所以$\lambda\leqslant\frac{n(n+1)}{2^{n-2}}$。设$f(n)=\frac{n(n+1)}{2^{n-2}}$，计算得$f(1)=4$，$f(2)=f(3)=6$，$f(4)=5$，$f(5)=\frac{15}{4}$，且$f(n+1)-f(n)=\frac{(n+1)(2-n)}{2^{n-1}}$，所以当$n\geqslant3$时，$f(n+1)<f(n)$。
因为M的元素个数为3，所以$\lambda\in(4,5]$。

Explanation

本题主要考查等差数列和等比数列。
（1）由于b_n为等差数列，则可根据已知解得b_1、d，从而得到数列b_n的通项公式；
（2）①由于$a_{10}=b_4=8$，又由已知得$a_{13}=1$，从而得到q，则可得到数列C_n的通项公式，可以看出C_n为等比数列和等差数列乘积的形式，利用错位相减法即可得到其前n项和S_n；
②由已知可得$\lambda\leqslant\frac{n(n+1)}{2^{n-2}}$，设$f(n)=\frac{n(n+1)}{2^{n-2}}$，根据$f(n)$的单调性即可得到$\lambda$的取值范围。

Model Response (Claude 3.5 Sonnet)

(1) 求数列$\{b_n\}$的通项公式
数列$\{b_n\}$是等差数列，设其公差为d。已知$b_2=a_2=4$，$b_5=a_5=10$。根据等差数列的性质，$b_5=b_2+3d$，所以$10=4+3d$，解得$d=2$。又因为$b_1=b_2-d=4-2=2$，所以数列$\{b_n\}$的通项公式为$b_n=b_1+(n-1)d=2+2(n-1)=2n$。
(2) ①求S_n
设每一行的公比为q。数列$\{C_n\}$的项分别是表中每一行的中间项，即$C_1=a_1$，$C_2=a_3$，$C_3=a_7$，... 由于从第二行起，每一行都是等比数列，且$a_{13}=1$。观察数表，$a_1=b_1=2$，$a_2=b_2=4$，$a_3=a_2\cdot q=4q$，$a_4=a_2\cdot q^2=4q^2$。$a_5=b_3=6$，$a_6=a_5\cdot q=6q$，$a_7=a_5\cdot q^2=6q^2$，$a_8=a_5\cdot q^3=6q^3$，$a_9=a_5\cdot q^4=6q^4$。a_{13}位于第4行，第4行有7个数，中间项为第4个，即$a_{13}=b_4\cdot q^3=8q^3=1$，所以$q^3=\frac{1}{8}$，$q=\frac{1}{2}$。因此，$C_n=b_n\cdot q^{n-1}=2n\cdot(\frac{1}{2})^{n-1}=n\cdot(\frac{1}{2})^{n-2}=4n\cdot(\frac{1}{2})^n$。$S_n=\sum_{k=1}^n C_k=\sum_{k=1}^n 4k(\frac{1}{2})^k=4\sum_{k=1}^n k(\frac{1}{2})^k$。利用错位相减法求和：$S=\sum_{k=1}^n kx^k=x+2x^2+3x^3+\ldots+nx^n$
$xS=x^2+2x^3+\ldots+(n-1)x^n+nx^{n+1}$ $(1-x)S=x+x^2+x^3+\ldots+x^n-nx^{n+1}=\frac{x(1-x^n)}{1-x}-nx^{n+1}$ $S=\frac{x(1-x^n)}{(1-x)^2}-\frac{nx^{n+1}}{1-x}$
当$x=\frac{1}{2}$时，$S_n=4[\frac{\frac{1}{2}(1-(\frac{1}{2})^n)}{(1-\frac{1}{2})^2}-\frac{n(\frac{1}{2})^{n+1}}{1-\frac{1}{2}}]=8[1-(\frac{1}{2})^n]-4n(\frac{1}{2})^n=8-(n+8)(\frac{1}{2})^n$。
(2) ②求λ的取值范围
$M=\{n|(n+1)C_n\geqslant\lambda, n\in\mathbb{N}^*\}$，$|M|=3$。$(n+1)C_n=(n+1)\cdot4n(\frac{1}{2})^n=4n(n+1)(\frac{1}{2})^n=f(n)$。$f(1)=8, f(2)=12, f(3)=12, f(4)=10, f(5)=7.5, f(6)=5.3...$ $f(n)$先增后减，且$f(2)=f(3)$。因为$|M|=3$，所以λ的取值范围为$7.5<\lambda\leqslant10$。
最终答案：
1. $b_n=2n$ 2. ①$S_n=8-(n+8)(\frac{1}{2})^n$ ②$7.5<\lambda\leqslant10$

Error Analysis

Error Type: Calculation Error
Analysis
The model exhibits significant **calculation errors** in Part (2) of its response, which propagate to affect the final answers. Firstly, when determining the sum S_n, after correctly setting up the arithmetic-geometric series summation $4\sum_{k=1}^n k(\frac{1}{2})^k$ and applying a general formula, the model's simplification $S_n=4\left[\frac{\frac{1}{2}(1-(\frac{1}{2})^n)}{(1-\frac{1}{2})^2}-\frac{n(\frac{1}{2})^{n+1}}{1-\frac{1}{2}}\right]$ results in $8-(n+8)(\frac{1}{2})^n$. The correct simplification of its own intermediate steps should be $8-(4n+8)(\frac{1}{2})^n$, which is equivalent to the standard solution's $S_n=8-\frac{n+2}{2^{n-2}}$. This discrepancy arises from an error in combining terms during the calculation. Secondly, when calculating $f(n)=(n+1)C_n=4n(n+1)(\frac{1}{2})^n$ for part (2)②, the model correctly states the formula for $f(n)$, but then lists numerical values $f(1)=8, f(2)=12, f(3)=12, f(4)=10, f(5)=7.5$. These listed values are exactly double the correct values that its own formula $f(n)=\frac{n(n+1)}{2^{n-2}}$ would produce ($f(1)=4, f(2)=6, f(3)=6, f(4)=5, f(5)=3.75$). These specific numerical miscalculations are central to the incorrect range derived for λ.

Figure 34: An example of Calculation Error.

Question & Standard Solution

Question
Problem (Q&A): (This question is worth 13 points) Arrange all terms of the sequence $\{a_n\}$ into a table where each row has two more terms than the previous row: $a_1, a_2 \quad a_3 \quad a_4, a_5 \quad a_6 \quad a_7 \quad a_8 \quad a_9, \cdots\cdots$. It is known that the first column of the table, $a_1, a_2, a_5, \cdots$, forms an arithmetic sequence, denoted as $\{b_n\}$, with $b_2 = 4$ and $b_5 = 10$. The numbers in the middle of each row, $a_1, a_3, a_7, \cdots$, form the sequence $\{C_n\}$, with the sum of its first n terms denoted as S_n. (1) Find the general formula for the sequence $\{b_n\}$; (2) If, starting from the second row, the numbers in each row from left to right form a geometric sequence with the same common ratio, and $a_{13} = 1$, find: ①S_n; ②Let $M = \{n|\ (n+1)C_n \geqslant \lambda, n \in \mathbb{N}^*\}$. If the set M has exactly 3 elements, find the range of λ.
Standard Solution
(1) Since $\{b_n\}$ is an arithmetic sequence with $b_2 = 4$ and $b_5 = 10$, let the common difference be d. Then, $b_1 + d = 4$ and $b_1 + 4d = 10$. Solving, we get $b_1 = 2$, $d = 2$. Thus, the general formula for the sequence $\{b_n\}$ is $b_n = 2n$.
(2) ①Let the common ratio of the geometric sequence in each row be q. The total number of terms in the first n rows is $1+3+5+\cdots+(2n-1) = n^2$. Since $3^2 < 13 < 4^2$, we have $a_{10} = b_4 = 8$, and $a_{13} = a_{10}q^3 = 8q^3 = 1$, so $q = \frac{1}{2}$. Therefore, $C_n = 2n \cdot \left(\frac{1}{2}\right)^{n-1} = \frac{n}{2^{n-2}}$.
Thus, $S_n = \frac{1}{2^{-1}} + \frac{2}{2^0} + \frac{3}{2^1} + \cdots + \frac{n-1}{2^{n-3}} + \frac{n}{2^{n-2}}$. Then, $\frac{1}{2}S_n = \frac{1}{2^0} + \frac{2}{2^1} + \frac{3}{2^2} + \cdots + \frac{n-1}{2^{n-2}} + \frac{n}{2^{n-1}}$. Subtracting, we get $\frac{1}{2}S_n = \frac{1}{2^{-1}} + \frac{1}{2^0} + \frac{1}{2^1} + \cdots + \frac{1}{2^{n-2}} - \frac{n}{2^{n-1}} = 4 - \frac{n+2}{2^{n-1}}$. Thus, $S_n = 8 - \frac{n+2}{2^{n-2}}$.
②From ①, we know $C_n = \frac{n}{2^{n-2}}$, so $\lambda \leqslant \frac{n(n+1)}{2^{n-2}}$. Let $f(n) = \frac{n(n+1)}{2^{n-2}}$. Calculating, we get $f(1) = 4$, $f(2) = f(3) = 6$, $f(4) = 5$, $f(5) = \frac{15}{4}$, and $f(n+1) - f(n) = \frac{(n+1)(2-n)}{2^{n-1}}$. Thus, when $n \geqslant 3$, $f(n+1) < f(n)$.
Since M has exactly 3 elements, $\lambda \in (4, 5]$.

Explanation

This question mainly tests arithmetic and geometric sequences.
(1) Since $\{b_n\}$ is an arithmetic sequence, we can use the given conditions to solve for b_1 and d, thereby obtaining the general formula for $\{b_n\}$;
(2) ①Since $a_{10} = b_4 = 8$ and it is given that $a_{13} = 1$, we can find q. Then, we can obtain the general formula for the sequence $\{C_n\}$, which is in the form of a product of an arithmetic and geometric sequence. Using the method of staggered subtraction, we can find the sum of the first n terms, S_n;
②From the given conditions, we have $\lambda \leqslant \frac{n(n+1)}{2^{n-2}}$. Let $f(n) = \frac{n(n+1)}{2^{n-2}}$. Based on the monotonicity of $f(n)$, we can determine the range of λ.

Model Response (Claude 3.5 Sonnet)

(1) Find the general formula for the sequence $\{b_n\}$
The sequence $\{b_n\}$ is an arithmetic sequence with common difference d. Given $b_2 = a_2 = 4$, $b_5 = a_5 = 10$. By the properties of an arithmetic sequence, $b_5 = b_2 + 3d$, so $10 = 4 + 3d$, solving gives $d = 2$. Also, $b_1 = b_2 - d = 4 - 2 = 2$, so the general formula for $\{b_n\}$ is $b_n = b_1 + (n-1)d = 2 + 2(n-1) = 2n$.
(2) ①Find S_n
Let the common ratio of each row be q. The terms of the sequence $\{C_n\}$ are the middle terms of each row, i.e., $C_1 = a_1$, $C_2 = a_3$, $C_3 = a_7$, ... Since, starting from the second row, each row forms a geometric sequence and $a_{13} = 1$, observe the table: $a_1 = b_1 = 2$, $a_2 = b_2 = 4$, $a_3 = a_2 \cdot q = 4q$, $a_4 = a_2 \cdot q^2 = 4q^2$. $a_5 = b_3 = 6$, $a_6 = a_5 \cdot q = 6q$, $a_7 = a_5 \cdot q^2 = 6q^2$, $a_8 = a_5 \cdot q^3 = 6q^3$, $a_9 = a_5 \cdot q^4 = 6q^4$. a_{13} is in the 4th row, which has 7 numbers, with the middle term being the 4th, i.e., $a_{13} = b_4 \cdot q^3 = 8q^3 = 1$, so $q^3 = \frac{1}{8}$, $q = \frac{1}{2}$. Thus, $C_n = b_n \cdot q^{n-1} = 2n \cdot \left(\frac{1}{2}\right)^{n-1} = n \cdot \left(\frac{1}{2}\right)^{n-2} = 4n \cdot \left(\frac{1}{2}\right)^n$. $S_n = \sum_{k=1}^n C_k = \sum_{k=1}^n 4k\left(\frac{1}{2}\right)^k = 4\sum_{k=1}^n k\left(\frac{1}{2}\right)^k$. Using the staggered subtraction method: $S = \sum_{k=1}^n kx^k = x + 2x^2 + 3x^3 + \ldots + nx^n$ $xS = x^2 + 2x^3 + \ldots + (n-1)x^n + nx^{n+1}$ $(1-x)S = x + x^2 + x^3 + \ldots + x^n - nx^{n+1} = \frac{x(1-x^n)}{1-x} - nx^{n+1}$ $S = \frac{x(1-x^n)}{(1-x)^2} - \frac{nx^{n+1}}{1-x}$ When $x = \frac{1}{2}$, $S_n = 4\left[\frac{\frac{1}{2}(1-\left(\frac{1}{2}\right)^n)}{\left(1-\frac{1}{2}\right)^2} - \frac{n\left(\frac{1}{2}\right)^{n+1}}{1-\frac{1}{2}}\right] = 8\left[1 - \left(\frac{1}{2}\right)^n\right] - 4n\left(\frac{1}{2}\right)^n = 8 - (n+8)\left(\frac{1}{2}\right)^n$.
(2) ②Find the range of λ
$M = \{n|(n+1)C_n \geqslant \lambda, n \in \mathbb{N}^*\}$, $|M| = 3$. $(n+1)C_n = (n+1) \cdot 4n\left(\frac{1}{2}\right)^n = 4n(n+1)\left(\frac{1}{2}\right)^n = f(n)$. $f(1) = 8$, $f(2) = 12$, $f(3) = 12$, $f(4) = 10$, $f(5) = 7.5$, $f(6) = 5.3$... $f(n)$ increases then decreases, and $f(2) = f(3)$. Since $|M| = 3$, the range of λ is $7.5 < \lambda \leqslant 10$.
Final Answer:
1. $b_n = 2n$ 2. ①$S_n = 8 - (n+8)\left(\frac{1}{2}\right)^n$ ②$7.5 < \lambda \leqslant 10$

Error Analysis

Error Type: Calculation Error
Analysis
The model exhibits significant **calculation errors** in Part (2) of its response, which propagate to affect the final answers. Firstly, when determining the sum S_n, after correctly setting up the arithmetic-geometric series summation $4\sum_{k=1}^n k(\frac{1}{2})^k$ and applying a general formula, the model's simplification $S_n = 4\left[\frac{\frac{1}{2}(1-(\frac{1}{2})^n)}{(1-\frac{1}{2})^2} - \frac{n(\frac{1}{2})^{n+1}}{1-\frac{1}{2}}\right]$ results in $8 - (n+8)(\frac{1}{2})^n$. The correct simplification of its own intermediate steps should be $8 - (4n+8)(\frac{1}{2})^n$, which is equivalent to the standard solution's $S_n = 8 - \frac{n+2}{2^{n-2}}$. This discrepancy arises from an error in combining terms during the calculation. Secondly, when calculating $f(n) = (n+1)C_n = 4n(n+1)(\frac{1}{2})^n$ for part (2)②, the model correctly states the formula for $f(n)$, but then lists numerical values $f(1) = 8$, $f(2) = 12$, $f(3) = 12$, $f(4) = 10$, $f(5) = 7.5$. These listed values are exactly double the correct values that its own formula $f(n) = \frac{n(n+1)}{2^{n-2}}$ would produce ($f(1) = 4$, $f(2) = 6$, $f(3) = 6$, $f(4) = 5$, $f(5) = 3.75$). These specific numerical miscalculations are central to the incorrect range derived for λ.

Figure 35: The English translation of the example of Calculation Error in Fig. 34.

Model	Visual Misinterpretation	Textual Misunderstanding	Reasoning Error	Integration Failure	Calculation Error	Total Cases
GPT-4o	7.5%(3)	2.5%(1)	67.5%(27)	7.5%(3)	15.0%(6)	40
Claude 3.5 Sonnet	5.0%(2)	2.5%(1)	82.5%(33)	0.0%(0)	10.0%(4)	40
Gemini 1.5 Pro 002	7.5%(3)	0.0%(0)	75.0%(30)	5.0%(2)	12.5%(5)	40
Llama-3.2-90B-Vision-Instruct	5.0%(2)	2.5%(1)	85.0%(34)	0.0%(0)	7.5%(3)	40
Qwen2-VL-72B-Instruct	5.0%(2)	0.0%(0)	80.0%(32)	2.5%(1)	12.5%(5)	40
Qwen2.5-Math-72B-Instruct	-(-)	0.0%(0)	100.0%(10)	-(-)	0.0%(0)	10
DeepSeekMath-7B-Instruct	-(-)	10.0%(1)	90.0%(9)	-(-)	0.0%(0)	10
o1	20.0%(2)	0.0%(0)	60.0%(6)	0.0%(0)	20.0%(2)	10
Claude 3.7 Sonnet	40.0%(4)	0.0%(0)	40.0%(4)	0.0%(0)	20.0%(2)	10
Overall Average	7.5%(18)	1.7%(4)	77.1%(185)	2.5%(6)	11.3%(27)	240

Table 9: Error distribution across different categories for evaluated models. Percentages are followed by absolute counts in parentheses. Categories with -(-) indicate an impossible combination of model and error type.