PhysReason: A Comprehensive Benchmark towards Physics-Based Reasoning

Xinyu Zhang[1,4], **Yuxuan Dong**[1,4], **Yanrui Wu**[1,4], **Jiaxing Huang**[1,4], **Chengyou Jia** [1,4], **Basura Fernando** [2], **Mike Zheng Shou** [3], **Lingling Zhang**[1,4] *, **Jun Liu**[1,5]

[1]School of Computer Science and Technology, Xi'an Jiaotong University
[2]IHPC, Agency for Science, Technology and Research, Singapore
[3]Show Lab, National University of Singapore
[4]Ministry of Education Key Laboratory of Intelligent Networks and Network Security, China
[5]Shaanxi Province Key Laboratory of Big Data Knowledge Engineering, China
zhang1393869716@stu.xjtu.edu.cn, {zhanglling,liukeen}@xjtu.edu.cn

Abstract

Large language models demonstrate remarkable capabilities across various domains, especially mathematics and logic reasoning. However, current evaluations overlook physics-based reasoning, a complex task requiring physics theorems and constraints. We present PhysReason, a 1,200-problem benchmark comprising knowledge-based (25%) and reasoning-based (75%) problems, where the latter are divided into three difficulty levels (easy, medium, hard). Notably, problems require an average of 8.1 solution steps, with hard problems requiring 15.6, reflecting the complexity of physics-based reasoning. We propose the Physics Solution Auto Scoring Framework, incorporating efficient answer-level and comprehensive step-level evaluations. Top-performing models like Deepseek-R1, Gemini-2.0-Flash-Thinking, and o3-mini-high achieve less than 60% on answer-level evaluation, with performance dropping from knowledge questions (75.11%) to hard problems (31.95%). Through step-level evaluation, we identify four key bottlenecks: Physics Theorem Application, Physics Process Understanding, Calculation, and Physics Condition Analysis. These findings position PhysReason as a novel and comprehensive benchmark for evaluating physics-based reasoning capabilities in large language models. Our code and data will be published at https://dxzxy12138.github.io/PhysReason/.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable performance across various domains, such as math (Lightman et al., 2024; Cobbe et al., 2021) and logical reasoning (Hendrycks et al.; Xu et al., 2025; Zhang et al., 2024). However, current evaluations often overlook physics-based reasoning, limiting their applications in scenarios such as robotics (Chow et al., 2025) and autonomous driving (Huang et al., 2023). This is because physics-based reasoning, integrating multiple theorems and physics constraints, is more closely aligned with practical applications than math and logical reasoning. Consequently, developing a comprehensive benchmark for evaluating LLMs' physics-based reasoning capabilities is crucial for discovering current limitations and guiding future improvements.

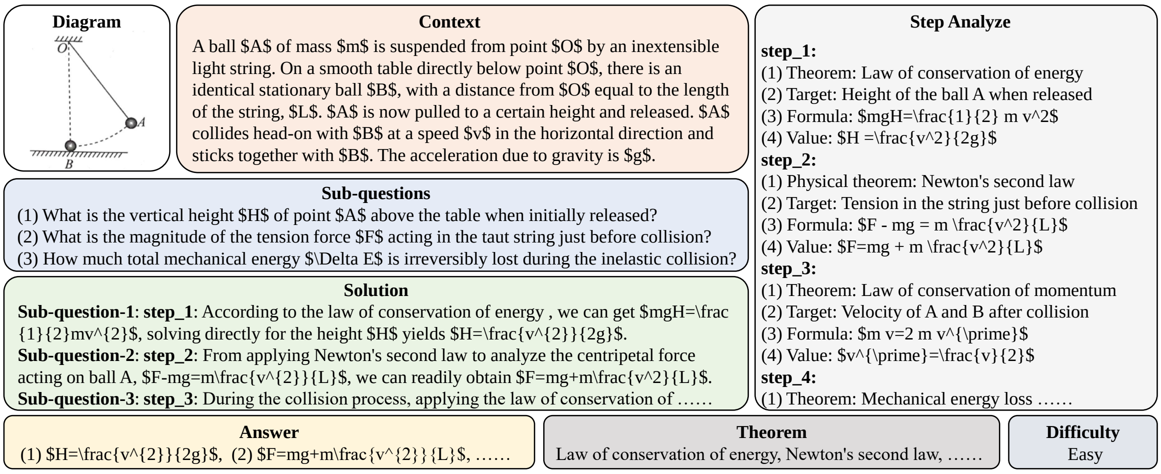
There are several pioneering physics benchmarks (K-12 level like ScienceQA (Lu et al., 2022), college-level SciBench (Wang et al.), and expert-level GPQA (Rein et al., 2024)) encompassing progressively advanced knowledge domains. However, they exhibit two critical limitations: oversimplified reasoning processes and neglecting step-level evaluation. These problems typically involve only 3-4 physics formulas, focusing solely on final answers to measure model performance. Therefore, a benchmark featuring in-depth reasoning processes and step-level evaluation is urgently needed to measure LLMs' physics-based reasoning capabilities.

To address these limitations, we present PhysReason, a comprehensive benchmark comprising 1,200 problems designed to evaluate models' physics-based reasoning capabilities. As illustrated in Figure 1, PhysReason features physics problems that require multi-step reasoning and precise application of physics theorems. The benchmark introduces several key characteristics:

1. **Stratified difficulty**: There are knowledge-based (25%) and reasoning-based (75%) problems, with reasoning problems categorized into easy, medium, and hard (25% each).
2. **Complex reasoning**: Solutions average 8.1 steps per problem, with hard problems reaching 15.6 steps, exceeding current physics benchmarks which typically only contain 3-4 steps.
3. **Multi-modal design**: 81% of problems include diagrams, evaluating models' capabilities in comprehending visual and textual information.

[*]Corresponding author

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16593–16615
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics



Context

A ball A of mass m is suspended from point O by an inextensible light string. On a smooth table directly below point O, there is an identical stationary ball B, with a distance from O equal to the length of the string, L. A is now pulled to a certain height and released. A collides head-on with B at a speed v in the horizontal direction and sticks together with B. The acceleration due to gravity is g.

Sub-questions

(1) What is the vertical height H of point A above the table when initially released?
(2) What is the magnitude of the tension force F acting in the taut string just before collision?
(3) How much total mechanical energy ΔE is irreversibly lost during the inelastic collision?

Solution

Sub-question-1: **step_1**: According to the law of conservation of energy , we can get $mgH=\frac {1}{2}mv^{2}$, solving directly for the height H yields $H=\frac{v^{2}}{2g}$.
Sub-question-2: **step_2**: From applying Newton's second law to analyze the centripetal force acting on ball A, $F-mg=m\frac{v^{2}}{L}$, we can readily obtain $F=mg+m\frac{v^2}{L}$.
Sub-question-3: **step_3**: During the collision process, applying the law of conservation of ……

Step Analyze

step_1:
(1) Theorem: Law of conservation of energy
(2) Target: Height of the ball A when released
(3) Formula: $mgH=\frac{1}{2} m v^2$
(4) Value: $H =\frac{v^2}{2g}$
step_2:
(1) Physical theorem: Newton's second law
(2) Target: Tension in the string just before collision
(3) Formula: $F - mg = m \frac{v^2}{L}$
(4) Value: $F=mg + m \frac{v^2}{L}$
step_3:
(1) Theorem: Law of conservation of momentum
(2) Target: Velocity of A and B after collision
(3) Formula: $m v=2 m v^{\prime}$
(4) Value: $v^{\prime}=\frac{v}{2}$
step_4:
(1) Theorem: Mechanical energy loss ……

Answer

(1) $H=\frac{v^{2}}{2g}$, (2) $F=mg+m\frac{v^{2}}{L}$, ……

Theorem

Law of conservation of energy, Newton's second law, ……

Difficulty

Easy

Figure 1: An illustration of example from our PhysReason benchmark. Due to space constraints, only key components are shown. Please refer to Appendix D for complete annotations.

To evaluate performance on PhysReason comprehensively, we propose the Physics Solution Auto Scoring Framework (PSAS) based on current LLMs' capabilities in information extraction and formula comparison. This framework encompasses two answer-level and step-level evaluation methods, PSAS-A and PSAS-S. PSAS-A enables efficient evaluation through answer comparison, while PSAS-S facilitates comprehensive analysis through step-by-step reasoning verification. Experimental results demonstrate that PSAS significantly outperforms direct LLM-based evaluation approaches, achieving an evaluation accuracy exceeding 98%.

We evaluate seven non-O-like models and eight O-like models on the PhysReason benchmark. Results show that while Deepseek-R1 (Guo et al., 2025), Gemini-2.0-Flash-Thinking-0121 (Deepmind), and o3-mini-high (OpenAI, 2025) demonstrate superior performance, their average scores remain below 60%. Moreover, models excel in basic physics concepts but consistently show performance degradation as problem difficulty and required solution steps increase (from 75.11% to 31.95%). This degradation stems from the models' inability to maintain accuracy across consecutive solution steps, so maintaining the reliability of the reasoning process is crucial. Through step-level evaluation, we identify four critical bottlenecks limiting model performance: Physics Theorem Application, Physics Process Understanding, Calculation Process, and Physics Condition Analysis.

2 Related Work

2.1 Large Language Model Evaluation

LLMs have demonstrated remarkable performance across various domains, including math reasoning (Jiang et al., 2024; Li et al., 2024; Imani et al., 2023), logical reasoning (Sun et al., 2024a; Xu et al., 2024; Zhang et al., 2025), and text generation (Zhao et al., 2024; Liang et al., 2024). However, these models exhibit notable limitations when confronted with physics-based interactions, significantly constraining their deployment in autonomous driving and robotics applications (Gao et al., 2024b). Unlike mathematical and logical reasoning, physics-based reasoning requires the sophisticated integration of multiple principles alongside real-world physical constraints (Kline, 1981). Consequently, developing robust physics reasoning capabilities represents a crucial prerequisite for expanding LLMs' potential in practical scenarios (Lai et al., 2024). Current evaluation methodologies predominantly focus on math or logical reasoning domains, revealing a critical gap in systematically assessing LLMs' physics reasoning proficiency.

2.2 Physics Benchmarks

Existing physics benchmarks span three knowledge complexity levels: K-12 (ScienceQA (Lu et al., 2022), E-EVAL (Hou et al., 2024)), college-level (MMLU (Hendrycks et al.), AGIEval (Zhong et al., 2024), JEEBench (Arora et al.), TheoremQA (Chen et al., 2023), EMMA (Hao et al., 2025), SciEval (Sun et al., 2024b), C-Eval-STEM (Huang et al., 2024), SciBench (Wang et al.)), and expert-level (OlympiadBench(He et al., 2024), GPQA (Rein et al., 2024)). While these benchmarks showcase LLMs' knowledge breadth, they simplify reasoning to 3-4 steps and emphasize only final answers. PhysReason addresses these gaps through complex reasoning process and step-level evaluation.

Table 1: Comparative analysis of our PhysReason with other physics-based reasoning benchmarks. For **Knowledge**, COMP: Competition, COL: College, CEE: College Entrance Examination, K1-K12: Elementary and High School, PH.D: Doctor of Philosophy; For **question type**, OE: Open-ended, MC: Multiple-choice, Avg. T: Average Tokens; For **solution type**, Avg. S: Average Steps.

Benchmark	Multi-modal	Size	Knowledge	Question		Solution		
				Type	Avg. T	Step-by-step	Avg. T	Avg. S
JEEBench	✗	123	CEE	OE,MC	169.7	-	-	-
MMLU-Pro	✗	1299	COL	MC	52.1	-	-	-
GPQA	✗	227	PH.D.	OE	111.4	✗	197.2	3.6
SciEval	✗	1657	-	OE,MC	154.5	-	-	-
SciBench	✓	295	COL	OE	80.5	✗	315.9	2.8
MMMU	✓	443	COL	OE,MC	53.8	-	-	-
ScienceQA	✓	617	K1-K12	MC	13.3	✗	63.0	2.4
OlympiadBench	✓	2334	COMP	OE	222.0	✗	199.8	3.7
EMMA	✓	156	-	MC	109.5	-	-	-
Ours-Knowledge	✓	300	CEE+COMP	OE	163.7	✓	196.5	3.3
Ours-Easy	✓	300	CEE+COMP	OE	171.2	✓	241.5	5.0
Ours-Medium	✓	300	CEE+COMP	OE	229.2	✓	391.3	8.4
Ours-Hard	✓	300	CEE+COMP	OE	340.9	✓	936.1	15.6
Ours-Full	✓	1200	CEE+COMP	OE	226.3	✓	441.3	8.1

3 Benchmark

3.1 Collection

We describe our comprehensive data collection process that spans five key stages: **Acquisition**, **Standardization**, **Translation**, **Search Prevention**, and **Difficulty Classification**.

Acquisition: We collect public physics problems from global college entrance examinations, their associated practice tests, and international physics competitions. Our sources include Chinese, Indian, and Russian exams, as well as IPhO, APhO, EPhO, and so on. This comprehensive benchmark derives from 1,254 PDFs containing over 20,000 unique problems, ensuring diverse difficulty levels.

Standardization: Using MinerU (Wang et al., 2024a) framework, we parse the content of these PDFs into structured problem information. Subsequently, all problems undergo rigorous deduplication, filtering, and formatting to ensure complete problem statements, precise physics terms, accurate expressions, and consistent presentation style.

Translation: We implement a two-phase translation process utilizing translators for initial conversion. Engineering Ph.D. candidates with physics expertise then verify the translations for accuracy and professionalism, especially physics terms.

Search Prevention: Following (Rein et al., 2024), we exclude problems whose solutions and answers can be found through a five-minute Google search to minimize potential data leakage.

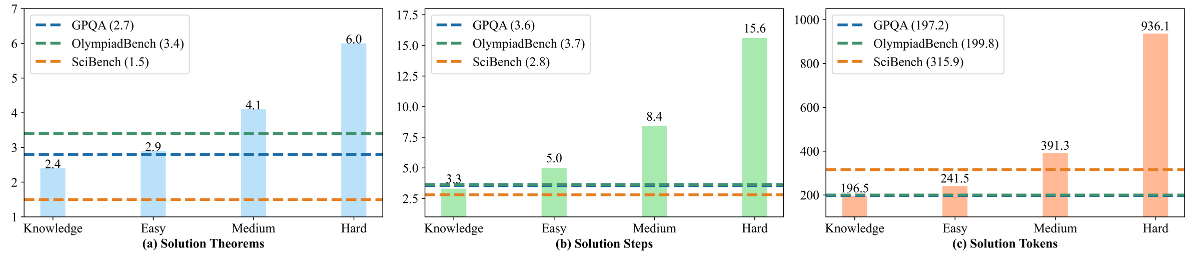
Difficulty Classification: Based on the time students typically need to solve problems and the theorems applied, questions are categorized into knowledge-based and reasoning-based types, with the latter subdivided into three difficulty levels (easy, medium, and hard). This classification enables the comprehension evaluation of physics concepts and physics-based reasoning capabilities.

3.2 Annotation

As shown in Figure 1, our annotation framework consists of 8 key elements: **Diagram**, **Context**, **Sub-questions**, **Solution**, **Step Analysis**, **Answer**, **Theorem**, and **Difficulty**. **Context** presents the physics scenario and conditions. **Diagram** visualizes the physics scenario with concise illustrations complementing the **Context**. **Sub-questions** give questions to assess the understanding and application of the concept. **Solution** provides a step-by-step reasoning process, and **Answer** gives the answer to each sub-question. **Step Analysis** explains the physics theorem used in each step and the physics quantities obtained. **Theorem** lists the physics theorems applied in the question, and **Difficulty** indicates the difficulty classification.

3.3 Characteristics

PhysReason consists of 1,200 carefully curated physics problems as shown in Table 1, with a strategic composition of 25% knowledge-based and 75% reasoning-based questions across various difficulty levels, collectively covering 147 physics theorems. The problems span Classic Mechanics, Quantum



(a) Solution Theorems **(b) Solution Steps** **(c) Solution Tokens**

Figure 2: Analysis of solution theorems, solution steps, and solution tokens across different problem categories, with comparisons from SciBench, GPQA, and OlympiadBench.

Mechanics, Fluid Mechanics, Thermodynamics, Electromagnetics, Optics, and Relativity. As shown in Figure 2, three critical solution metrics (theorem, step, and token) correlate positively with problem difficulty levels, validating the rationality of our difficulty classification. Notably, the medium and hard problems demonstrate higher complexity compared to existing benchmarks. PhysReason distinguishes itself through three key characteristics:

1. **Stratified difficulty**: The benchmark maintains a carefully balanced composition of knowledge-based (25%) and reasoning-based (75%) problems. The reasoning-based problems are methodically distributed across three difficulty levels (easy, medium, and hard - 25% each), enabling comprehensive capability evaluation.
2. **Complex reasoning**: Detailed step-by-step solution annotations accompany each problem. These annotated solutions demonstrate complex reasoning chains averaging 8.1 steps per problem, with hard problems requiring up to 15.6 steps, significantly surpassing the complexity of existing physics-based reasoning benchmarks.
3. **Multi-modal design**: The benchmark features a high proportion (81%) of problems with diagrams, authentically replicating physics-based reasoning scenarios while effectively evaluating both textual and visual reasoning capabilities.

4 Evaluation Framework

4.1 Why LLMs Can Evaluate?

Unlike multiple-choice problems, PhysReason contains open-ended answers and steps with diverse expressions but consistent semantics. Given that LLMs have demonstrated exceptional capabilities in both precise content extraction and formula consistency evaluation (Contributors, 2023; Gao et al., 2024a), they serve as practical tools for automated physics solution evaluation. Therefore, we propose automated answer-level and step-level evaluations, achieving comprehensive evaluation and avoiding labor-intensive manual assessment.

4.2 How Answer-Level Evaluation Works?

We develop Physics Solution Auto Scoring Framework-Answer Level (PSAS-A), which evaluates based on sub-question answers. Given a model's reasoning process M for a problem with sub-questions $\{q_1, q_2, \ldots, q_n\}$, we first extract the model's answers $\hat{a}_i$ for each q_i from M with an LLM. Then, we employ the LLM to verify if $\hat{a}_i$ is semantically consistent with the standard answer a_i of sub-question q_i. The comparison function $C(\hat{a}_i, a_i)$ returns 1 if consistent and 0 otherwise. Considering that the sub-questions with different steps should not carry equal weights in scoring, we use the length of annotation solution s_i of sub-question q_i, i.e., $|(s_i)|$ as a weighting scalar. The model's reasoning process M's answer-level score for each problem is calculated as follows:

$$\text{Score}(M) = \frac{\sum_{q_i} |(s_i)| \times C(\hat{a}_i, a_i)}{\sum_{q_i} |(s_i)|} \qquad (1)$$

4.3 What is the Step?

Considering the correct execution of the reasoning process, each step should satisfy the following three conditions: **Completeness**: Each step should contain a complete unit of logical reasoning **Independence**: Each step should be understandable and evaluable as an independent logical unit **Progression**: Each step should provide substantial progress in the problem-solving process, moving the solution forward Therefore, we define that each step in the annotation must include a formula derived from applying a physics-based theorem and its related calculations. Through physics theorems and formulas, we ensure that our defined steps maintain completeness, independence, and progression in physical reasoning, providing a solid foundation for subsequent content.

4.4 How Step-Level Evaluation Works?

The current mainstream evaluation approach (He et al., 2024) with LLMs relies on answers, failing to

Algorithm 1 Physics Solution Auto Scoring Framework-Step Level (PSAS-S)

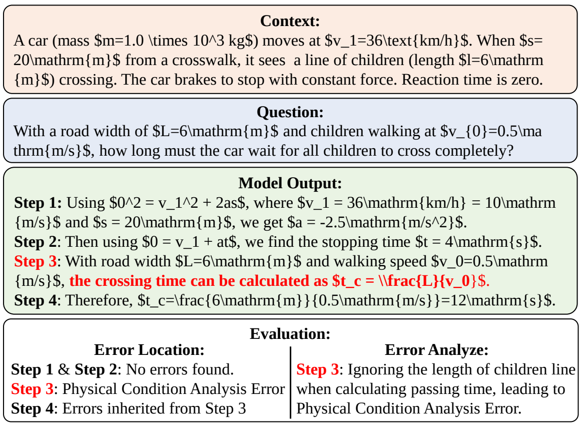
```
Phase 1: Data Extraction                                   ▷ Extract and normalize solution steps from model output
Input: Model output $M$, Annotation solution steps $S = \{s_1, s_2, ..., s_n\}$, Annotation step formulas $F = \{f_1, f_2, ..., f_n\}$,
Annotation step values $V = \{v_1, v_2, ..., v_n\}$        ▷ Information needed for step-level evaluation
for $s_i \in S$ do
    $E_i \leftarrow$ LLM(ExtractTemplate($M, s_i$))        ▷ E: extracted relevant steps
end for
Assert $|E| = |S|$                                         ▷ Ensure one-to-one mapping between extracted and annotated steps
Phase 2: Scoring                                           ▷ Evaluate formula application and numerical calculations
for $(e_i, s_i) \in (E, S)$ do
    $\hat{f}_i \leftarrow$ ExtractFormula($e_i$)           ▷ Formula content
    $\hat{v}_i \leftarrow$ ExtractValue($e_i$)             ▷ Calculation target
    $score_i \leftarrow 0.5 \times$ ScoreFormula($\hat{f}_i, f_i$) $+ 0.5 \times$ ScoreValue($\hat{v}_i, v_i$)   ▷ Get final score
end for
$final\_score \leftarrow \frac{\sum_{i=1}^{n} score_i}{n}$   ▷ Get the final score with the step-level evaluation
Phase 3: First Error Step Detection                        ▷ Identify the earliest point of solution deviation
$first\_error\_step \leftarrow \infty$                      ▷ Initialization
for $i \leftarrow 1$ to $|S|$ do
    if $score_i < 1$ then
        $error\_step \leftarrow$ FindOriStep($M, e_i$) with the relationship between $E$ and $M$   ▷ Find corresponding original step
        $first\_error\_step \leftarrow$ min($first\_error\_step, error\_step$)   ▷ Get the minimum
    end if
end for
Phase 4: Error Analysis                                    ▷ Analyze the first error step
ErrorTypes $\mathcal{T} = \{$DAE, PTAE, PCAE, PPUE, VRE, CPE, BCAE$\}$   ▷ Error categories
if $first\_error\_step < \infty$ then
    $j \leftarrow first\_error\_step$
    $error\_type \leftarrow$ LLM(ClassificationTemplate($e_j, s_j, \mathcal{T}$))   ▷ Identify error type
    $error\_analysis \leftarrow$ LLM(AnalysisTemplate($e_j, s_j$))   ▷ Generate error analysis
end if
Output: $final\_score, first\_error\_step, error\_type, error\_analysis$
```

reveal how and where models deviate from correct reasoning paths. To address this, we propose the Physics Solution Auto-Scoring Framework-Step Level (PSAS-S), which enables detailed assessment and analysis of each reasoning step. The framework is divided into four phases: **Data Extraction**, **Scoring**, **First Error Step Detection**, **Error Analysis**, as detailed in Algorithm 1.

Data Extraction phase leverages LLMs using *Target* components from *Step Analysis* annotations (Figure 1) as prompts to locate and extract relevant content from model outputs for each annotated solution step s_i. This phase effectively handles redundant thinking processes in LLM's reasoning process while maintaining semantic equivalence. It obtains the mapping relationship between extracted relevant steps E and annotated solution steps S.

Scoring phase evaluates each step s_i through two complementary components of theorem assessment $ScoreFormula(\hat{f}_i, f_i)$ and result verification $ScoreValue(\hat{v}_i, v_i)$, each with a weight of 0.5. The final score is calculated as shown in Algorithm 1. This ensures a balanced assessment of theorem application and computational accuracy.

First Error Step Detection phase identifies the earliest step of deviation from the correct solution path. When any step is found with a score below 1, FindOriStep function locates the corresponding original step in the model's raw output based on the mapping relationship between E and S obtained from the **Data Extraction** phase, and updates $first\_error\_step$ to maintain the earliest error position. This enables precise identification of where the model's reasoning first goes wrong.



```
Context:
A car (mass $m=1.0 \times 10^3 kg$) moves at $v_1=36\text{km/h}$. When $s=
20\mathrm{m}$ from a crosswalk, it sees  a line of children (length $l=6\mathrm
{m}$) crossing. The car brakes to stop with constant force. Reaction time is zero.

Question:
With a road width of $L=6\mathrm{m}$ and children walking at $v_{0}=0.5\ma
thrm{m/s}$, how long must the car wait for all children to cross completely?

Model Output:
Step 1: Using $0^2 = v_1^2 + 2as$, where $v_1 = 36\mathrm{km/h} = 10\mathrm
{m/s}$ and $s = 20\mathrm{m}$, we get $a = -2.5\mathrm{m/s^2}$.
Step 2: Then using $0 = v_1 + at$, we find the stopping time $t = 4\mathrm{s}$.
Step 3: With road width $L=6\mathrm{m}$ and walking speed $v_0=0.5\mathrm
{m/s}$, the crossing time can be calculated as $t_c = \\frac{L}{v_0}$.
Step 4: Therefore, $t_c=\frac{6\mathrm{m}}{0.5\mathrm{m/s}}=12\mathrm{s}$.

Evaluation:
Error Location:
Step 1 & Step 2: No errors found.
Step 3: Physical Condition Analysis Error
Step 4: Errors inherited from Step 3

Error Analyze:
Step 3: Ignoring the length of children line
when calculating passing time, leading to
Physical Condition Analysis Error.
```

Figure 3: Step-level evaluation example obtained from PSAS-S framework.

Error Analysis phase analyzes the first error step detected in the solution, with two components: error classification and error analysis. For error classification, PSAS-S considers seven types of common errors: Diagram Analysis Er-

Table 2: Comparison between PSAS framework and direct use of LLM evaluation, where Answer Acc. denotes the accuracy of answer-level evaluation and Step Acc. indicates the precision in identifying the initial error step in the reasoning process.

Model	Answer Acc.	Step Acc.
Gemini-2.0-Flash	87.81	33.18
Deepseek-V3	89.78	34.45
Gemini-2.0-Flash-Thinking-0121	91.24	35.74
Deepseek-R1	93.31	37.54
Our (Gemini-2.0-Flash)	98.96	97.23
Our (Deepseek-V3)	99.35	98.04

ror (DAE), Physics Theorem Application Error (PTAE), Physics Condition Analysis Error (PCAE), Physics Process Understanding Error (PPUE), Variable Relationship Error (VRE), Calculation Process Error (CPE), and Boundary Condition Analysis Error (BCAE). Detailed error-type descriptions are available in the Appendix C. LLMs use structured prompts to identify the error type for the first error step. Then, a comprehensive error analysis is generated to explain the reasoning behind the mistake. A simplified example is shown in Figure 3.

4.5 Whether Evaluation Trustworthy?

To validate the reliability of both our PSAS-A and PSAS-S, we compare our PSAS against conventional direct LLM evaluation approaches at both answer-level and step-level, using the Chain-of-Thought reasoning strategy. We implement experiments using Deepseek-V3 and Gemini-2.0-Flash as scoring models in the following experiments:

1. For answer-level evaluation, we employ scoring models to assess answer correctness by combining both model-generated outputs and annotation answers. We then compare these results with the judgments obtained from PSAS-A.
2. For step-level evaluation, inspired by previous work (Zheng et al., 2024), we design the task of identifying the first error step in reasoning processes containing errors, where higher accuracy indicates a more precise evaluation of the reasoning process. Then, we submit both model-generated and annotated reasoning processes to scoring models to determine the location of the first error step, comparing with PSAS-S.

Then, we collect 8,400 reasoning processes generated from multiple advanced models, including Deepseek-R1, Gemini-2.0-Flash, Gemini-2.0-Flash-Thinking-0121, GLM-Zero, o1-mini, QwQ-32B, and QvQ-72B. Subsequently, we randomly sample 1,000 reasoning processes and meticulously manually annotate them to determine the correctness of each answer and identify the location of the first error step. The results presented in Table 2 demonstrate that our frameworks achieve superior performance compared to direct LLM evaluation, highlighting the accuracy and reliability of PSAS evaluation results on PhysReason.

5 Experiments

5.1 Setting

Baselines: We evaluate current mainstream open-source and closed-source LLMs, VLMs, and several o-like models. For models that cannot accept visual inputs, we use Gemini-2.0-Flash to generate captions for each image as supplementary information. We assess 15 advanced LLMs/VLMs under the zero-shot Chain-of-Thought (CoT) setting (encouraging models to think step by step), including 7 non-O-like models (Qwen2-VL-72B (Wang et al., 2024b), GPT-4o (OpenAI), Claude-3.5-Sonnet (Anthropic), InternVL2.5-78B (Chen et al., 2024), Deepseek-v3 (DeepSeek-AI, 2024)), Gemini-2.0-Flash (Deepmind, 2024), Gemini-2.0-Pro (Deepmind, 2025) and 8 O-like models (QwQ-32B (Team, 2024b), QvQ-72B (Team, 2024a), o1-mini (OpenAI, 2024b), o1 (OpenAI, 2024a), o3-mini-high (OpenAI, 2025), Gemini-2.0-Flash-Thinking (Deepmind), Deepseek-R1 (Guo et al., 2025), GLM-Zero (ZhipuAI, 2024)). Note that Gemini-2.0-Flash-Thinking has two versions: 1206 and 0121. Due to API limitations, we do not experiment with o1 on the entire dataset. All other models are evaluated on the complete benchmark.

Evaluation Workflow: We encourage models to generate reasoning processes step by step for all problems in PhysReason, with open-source models running on NVIDIA A800 GPUs. Please refer to Appendix-E for the detail prompt template. Then, we evaluate the models' performance with the PSAS framework at both the answer and step levels, as described in Sections 4.2 and 4.4. Based on the experimental results in Section 4.5, considering both efficiency and performance, we select Deepseek-V3 as the final scoring model.

PhysReason-mini: Considering that the complete PhysReason requires relative high evaluation costs, we create a balanced PhysReason subset - PhysReason-mini. We randomly sample 200 questions from the whole benchmark (50 for each difficulty level), striving to achieve equal representation across categories wherever possible.

Table 3: Model performance comparisons on the PhysReason benchmark using answer-level (left of /) and step-level (right of /) evaluations across different input combinations of Questions (Q), Images (I), and Image Captions (IC). Gemini-2.0-T† and * represent Gemini-2.0-Flash-Thinking-1206 and 0121.

Model	Input	Knowledge	Easy	Medium	Hard	Avg.
Non-O-like Models						
Qwen2VL-72B	Q, I	41.92/62.47	24.04/45.26	15.97/36.13	4.83/24.23	16.96/42.88
InternVL2.5-78B	Q, I	28.34/64.71	24.16/50.69	17.72/38.56	9.71/25.95	19.98/45.89
GPT-4o	Q, I	50.71/65.82	33.87/51.98	22.73/42.36	11.03/24.71	29.58/47.23
Deepseek-V3-671B	Q, IC	55.86/66.14	40.06/52.77	26.63/44.02	13.73/26.87	34.07/48.42
Claude-3.5-Sonnet	Q, I	54.14/66.45	41.35/55.85	28.14/44.86	15.11/28.51	34.69/49.88
Gemini-2.0-Flash	Q, I	65.08/75.04	54.84/68.60	39.79/55.67	21.99/38.39	45.20/60.40
Gemini-2.0-Pro	Q, I	67.99/79.01	55.43/71.47	44.29/57.74	23.81/42.66	47.88/62.74
O-like Models						
o1-mini	Q, IC	53.90/65.74	35.21/52.26	22.24/40.19	10.61/26.80	30.49/47.18
QvQ-72B	Q, I	62.44/70.92	53.74/64.65	28.18/54.88	14.30/36.47	32.67/57.66
Gemini-2.0-T†	Q, I	65.35/77.20	51.89/67.49	44.43/58.95	27.14/45.48	47.20/63.07
QwQ-32B	Q, IC	62.03/76.28	54.92/71.08	43.64/62.14	22.99/42.19	45.89/63.87
GLM-Zero	Q, IC	64.95/80.36	54.11/71.54	41.32/63.67	23.04/47.46	46.52/65.76
o3-mini-high	Q, IC	70.67/83.61	67.20/81.95	45.31/64.57	30.12/47.23	53.32/69.34
Gemini-2.0-T*	Q, I	73.44/84.15	63.17/75.94	50.41/66.60	31.90/48.47	54.73/69.73
Deepseek-R1	Q, IC	75.11/85.91	65.08/79.81	54.84/72.02	31.95/51.50	56.75/73.26

Table 4: Comparison on PhysReason-mini with PSAS-A, where Gemini-2.0-T† and * represent Gemini-2.0-Flash-Thinking-1206 and 0121. And K., E., M. and H. represent knowledge, easy, medium and hard.

Model	K.	E.	M.	H.	Avg.
o1-mini	54.80	30.33	15.41	7.92	27.11
QvQ-72B	51.17	37.10	29.83	22.13	35.06
QwQ-32B	64.40	50.07	38.88	27.45	45.20
Gemini-2.0-T†	71.47	49.97	36.83	22.97	45.42
GLM-Zero	72.70	50.17	43.42	24.70	47.75
o1	72.47	53.37	49.31	25.32	50.12
o3-mini-high	71.10	63.20	47.02	31.93	53.31
Gemini-2.0-T*	76.33	56.87	51.85	32.61	54.42
Deepseek-R1	85.17	60.77	47.24	33.23	56.60

5.2 Main Results

As demonstrated in Tables 3 and 4, the experimental results on the PhysReason and PhysReason-mini reveal several significant findings.

Model Categories: O-like models exceed non-O-like ones, with multiple O-like models surpassing 50% answer-level accuracy compared to non-O-like models' peak of 47.88%.

Difficulty Level Analysis: As the difficulty increases, the required solution steps also increase, while model performance severely declines, indicating that models still perform inadequately on physics problems requiring deep reasoning.

Step-level vs. Answer-level Evaluation: The two evaluation frameworks assess performance from different perspectives. Step-level scores consistently surpass answer-level scores, indicating that models can achieve some correct steps despite failing to reach the correct final answer. Moreover, the step-level score differences between models become more pronounced than those at the answer level as problem difficulty increases. This demonstrates that step-level evaluation proves more discriminative in distinguishing model capabilities, particularly in highly challenging problems. The distributions of these two evaluation methodologies exhibit non-perfect synchronization, indicating that step-level evaluation provides comprehensive insights to answer-level assessment.

Medium and Hard Problem Analysis: Performance on medium and hard reasoning problems can emerge as key differentiators of model physics-based reasoning ability. Among these models, those achieving scores of 40/60 and 30/50 on answer-level and step-level evaluations respectively serve as critical reference points.

Knowledge-Reasoning Correlation Analysis: Results show a positive correlation between physics knowledge and reasoning capabilities, with Deepseek-R1 and Gemini-2.0-Flash-Thinking-0121 excelling in both aspects. Moreover, among models with similar scores on knowledge problems, O-like models tend to achieve higher scores on reasoning problems (as demonstrated by Gemini-2.0-Flash and Gemini-2.0-T†). This suggests that reinforcement learning and training with thought chains help improve models' reasoning capabili-

Table 5: Test-Time Compute Scaling Performance Comparisons on PhysReason-mini with PSAS-A, where Flash and Think denote Gemini-2.0-Flash and Gemini-2.0-Flash-Thinking-0121, and Tour. means Tournament.

Base	Method	Reward	N=1	N=2	N=4	N=8
Flash	BoN	Flash	46.52	46.67	47.12	47.81
		Think	46.52	47.37	48.87	50.94
	Tour.	Flash	46.52	45.87	47.36	49.58
		Think	46.52	47.51	52.11	53.06
Think	BoN	Think	54.42	52.27	54.78	55.13
	Tour.	Think	54.42	55.60	56.26	56.57

Table 6: Performance Comparison with PSAS-A after Directly Concatenation (D. Acc) and Guided Error Localization (G. Acc) on PhysReason-mini, where Acc. means the original performance of the model, Gemini-2.0-T* represents Gemini-2.0-Flash-Thinking-0121.

Model	Acc.	D. Acc.	G. Acc.
Deepseek-V3	34.07	29.31	40.78
Gemini-2.0-Flash	46.52	42.76	51.55
Gemini-2.0-T*	54.42	50.66	56.82
Deepseek-R1	56.60	52.26	58.33

ties. In conclusion, effective reasoning relies on knowledge capacity and model architecture.

5.3 Results with Test-Time Compute Scaling

We evaluate Best-of-N (BoN) and Tournament-Style selection (Snell et al., 2025; Yang et al., 2024) test-time compute scaling methods on PhysReason-mini. Using Gemini-2.0-Flash and Gemini-2.0-Flash-Thinking-0121 as base models, we test different reward model configurations: when Flash serves as base model, both itself and Thinking-0121 are evaluated as reward models, while Thinking-0121 uses self-reward due to its superior reasoning. Both methods (Cobbe et al., 2021; Lightman et al., 2024; Son et al., 2024) select optimal responses from multiple Chain-of-Thought candidates (N = 1, 2, 4, 8), as shown in Table 5. These scaling methods demonstrate the potential to enhance model performance through strategic response selection and process reward modeling.

5.4 Performance Improving with PSAS-S

Given PSAS-S's capability to locate and analyze the first error step as presented in Section 4.4, we conduct experiments on PhysReason-mini to explore whether models can correct errors after being informed. The experiments are divided into *Direct concatenation* and *Guided error localization*. The former (D. Acc.) combines questions with the previous reasoning process for a second attempt. For the latter (G. Acc.), PSAS-S is used to locate and analyze the first error in the reasoning process, then combines the question, previous reasoning, and **the location and analysis of the first error** for a second attempt. As shown in Table 6, results show that direct concatenation decreased performance by 3-5%, while guided error localization improved performance by 3-6%. This suggests that guiding LLMs to identify reasoning errors is crucial for enhancing their reasoning capabilities and also proves the effectiveness of our PSAS framework.

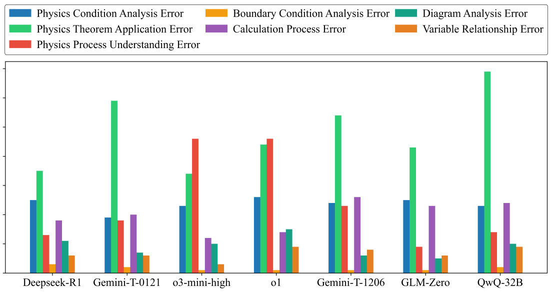


Figure 4: Error statistics with PSAS-S framwork in PhysReason-mini, where Gemini-T-1206 and Gemini-T-0121 denote Gemini-2.0-Flash-Thinking-1206 and Gemini-2.0-Flash-Thinking-0121.

5.5 Error Kind Distribution Analysis

Discovering errors in reasoning processes is not equivalent to fully understanding them; it's also crucial to understand the causes of errors. We analyze the error distributions of different models on PhysiReason-mini as shown in Figure 4. Four prevalent error types consistently challenge all models: Physics Theorem Application, Physics Process Understanding, Calculation Process, and Physics Condition Analysis. This reveals models' limited intuitive physics understanding, highlighting the need for stronger physics-based reasoning capabilities. Notably, o1 and o3-mini-high show elevated Physics Process Understanding Errors but reduced Calculation Process Errors. This maybe suggest a trade-off between conceptual comprehension and computational precision.

5.6 Hard Problem Analysis

Our analysis of 50 hard reasoning problems from PhysReason-mini across 7 models reveals two key insights (Figure 5). Despite variations in overall performance, each model exhibits unique strengths in specific problem domains, demonstrating the diverse nature of their reasoning capabilities. The models' achievement of some scores (below 1) is notable, indicating their ability to initiate correct

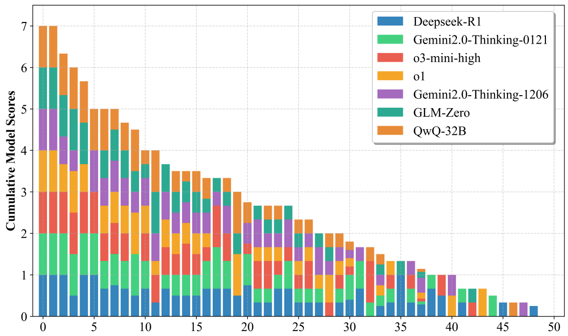


Figure 5: Performance with PSAS-S framework in the hard problems from PhysReason-mini.

solution paths but failing to maintain this accuracy throughout the reasoning process. These patterns suggest that while current models grasp basic physics concepts, they struggle to sustain accurate reasoning across extended solution steps.

6 Conclusion

We introduce PhysReason, a novel physics-based reasoning benchmark with stratified difficulty and Physics Solution Auto-Scoring Framework with answer and step level evaluation. Experimental results show a consistent decline in performance as reasoning depth increases. This benchmark establishes new standards for evaluating and improving AI models' physics-based reasoning abilities.

7 Acknowledgements

This work was supported by the National Key Research and Development Program of China (2022YFC3303600), National Natural Science Foundation of China (No. 62137002, 62293550, 62293553, 62293554, 62450005, 62437002, 62477036, 62477037, 62176209, 62192781, 62306229), 'LENOVO-XJTU' Intelligent Industry Joint Laboratory Project, the Shaanxi Provincial Social Science Foundation Project (No. 2024P041), the Natural Science Basic Research Program of Shaanxi (No. 2023-JC-YB-593), the National Research Foundation, Singapore, under its NRF Fellowship (Award# NRF-NRFF14-2022-0001), the Youth Innovation Team of Shaanxi Universities 'Multi-modal Data Mining and Fusion', Project of China Knowledge Centre for Engineering Science and Technology, and the Youth AI Talents Fund of China Association of Automation (Grant No. HBRC-JKYZD-2024-311). Mike Shou does not receive any funding for this work.

Limitation

Despite the comprehensive nature of our benchmark, two key limitations warrant discussion, concerning both benchmark construction and evaluation methodology. First, they focus primarily on testing models' ability to apply and reason with physics theorems under idealized conditions, rather than fully reflecting real-world physics scenarios. However, it is worth noting that applying physics theorems under idealized conditions serves as the foundation for real-world physics scenarios, as the latter is more complex. However, current LLMs' performance even on idealized conditions remains unsatisfactory. Therefore, PhysReason remains valuable in evaluating models' ability to apply physics theorems for physics-based reasoning. Moreover, through data synthesis, many problems in PhysReason can be adapted to create real-world physics reasoning scenarios, which will be a direction for our future research. Second, our evaluation framework, though achieving over 98% accuracy using LLMs as assessment tools, is not without limitations. The PSAS-S framework, while demonstrating satisfactory performance, increases computational time for evaluation. In future work, we will explore ways to optimize evaluation time while maintaining assessment accuracy.

Ethical Statement

In developing PhysReason, we carefully considered and addressed potential implications and risks. Our benchmark, sourced exclusively from public official materials (IPhO, Gaokao, JEE, and authorized mock exams), undergoes rigorous data cleansing, deduplication, and standardization to ensure reliability while minimizing bias and data leakage. Committed to environmental sustainability, we publicly release complete datasets and accompanying scripts under appropriate licenses (MIT and CC BY-NC-SA) to cut down on unnecessary carbon footprint, while optimizing processing pipelines to reduce computational overhead. In all experiments, we strictly comply with all licenses for models and data. Our benchmark is an important resource that drives AGI's strength in scientific reasoning, maintaining high standards for data quality and ethical considerations.

References

Anthropic. Claude 3.5 sonnet.

Daman Arora, Himanshu Gaurav Singh, et al. Have llms advanced enough? a challenging problem solving benchmark for large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

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

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

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. 2025. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*.

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

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`.

Google Deepmind. Gemini 2.0 flash thinking mode.

Google Deepmind. 2024. Introducing gemini 2.0: Our new ai model for the agentic era.

Google Deepmind. 2025. Introducing gemini 2.0 pro.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, et al. 2024a. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*.

Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. 2024b. Physically grounded vision-language models for robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12462–12469. IEEE.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. 2025. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Jinchang Hou, Chang Ao, Haihong Wu, Xiangtao Kong, Zhigang Zheng, Daijia Tang, Chengming Li, Xiping Hu, Ruifeng Xu, Shiwen Ni, and Min Yang. 2024. E-EVAL: A comprehensive Chinese k-12 education evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7753–7774, Bangkok, Thailand. Association for Computational Linguistics.

Yu Huang, Yue Chen, and Zhu Li. 2023. Applications of large scale foundation models for autonomous driving. *arXiv preprint arXiv:2311.12144*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 37–42.

Weisen Jiang, Han Shi, Longhui Yu, Zhengying Liu, Yu Zhang, Zhenguo Li, and James Kwok. 2024. Forward-backward reasoning in large language models for mathematical verification. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6647–6661.

Morris Kline. 1981. *Mathematics and the physical world*. Courier Corporation.

Wenqiang Lai, Tianwei Zhang, Tin Lun Lam, and Yuan Gao. 2024. Vision-language model-based physical reasoning for robot liquid perception. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9652–9659. IEEE.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. 2024. Snapkv: Llm knows what you are looking for before generation. *arXiv preprint arXiv:2404.14469*.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, et al. 2024. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering.

OpenAI. Hello gpt-4o.

OpenAI. 2024a. Learning to reason with LLMs.

OpenAI. 2024b. Openai o1-mini.

OpenAI. 2025. Openai o3-mini.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling llm test-time compute optimally can be more effective than scaling model parameters. In *International Conference on Learning Representations*.

Seonil Son, Ju-Min Oh, Heegon Jin, Cheolhun Jang, Jeongbeom Jeong, and Kuntae Kim. 2024. Varco arena: A tournament approach to reference-free benchmarking large language models. *arXiv preprint arXiv:2411.01281*.

Hongda Sun, Weikai Xu, Wei Liu, Jian Luan, Bin Wang, Shuo Shang, Ji-Rong Wen, and Rui Yan. 2024a. Determlr: Augmenting llm-based logical reasoning from indeterminacy to determinacy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9828–9862.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024b. 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.

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

Qwen Team. 2024b. Qwq: Reflect deeply on the boundaries of the unknown.

Zaharov Timur, Konstantin Korolev, and Aleksandr Nikolich. 2024. Physics big.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024a. Mineru: An open-source solution for precise document content extraction. *Preprint*, arXiv:2409.18839.

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

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*.

Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2025. Are large language models really good logical reasoners? a comprehensive evaluation and beyond. *IEEE Transactions on Knowledge and Data Engineering*.

Fangzhi Xu, Zhiyong Wu, Qiushi Sun, Siyu Ren, Fei Yuan, Shuai Yuan, Qika Lin, Yu Qiao, and Jun Liu. 2024. Symbol-LLM: Towards foundational symbol-centric interface for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13091–13116, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Lingling Zhang, Yifei Li, Qianying Wang, Yun Wang, Hang Yan, Jiaxin Wang, and Jun Liu. 2024. Fprompt-plm: Flexible-prompt on pretrained language model for continual few-shot relation extraction. *IEEE Transactions on Knowledge and Data Engineering*, 36(12):8267–8282.

Lingling Zhang, Yujie Zhong, Qinghua Zheng, Jun Liu, Qianying Wang, Jiaxin Wang, and Xiaojun Chang. 2025. Tdgi: Translation-guided double-graph inference for document-level relation extraction. *IEEE Transactions on Pattern Analysis & Machine Intelligence*.

Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru

Tang, Rui Zhang, and Arman Cohan. 2024. Docmath-eval: Evaluating math reasoning capabilities of llms in understanding long and specialized documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16103–16120.

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

ZhipuAI. 2024. Glm-zero mode.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314.

A Data Sources

Our dataset is derived from four distinct sources, each representing different levels and approaches to physics education and assessment. These sources have been carefully selected to ensure comprehensive coverage of physics problems across various difficulty levels and cultural contexts. The diversity of these sources helps in creating a robust and well-rounded dataset that captures different pedagogical approaches and problem-solving methodologies.

- **International Physics Olympiad (IPhO) Problems**
  The Physics Olympiad problems are globally recognized for their complexity and quality. These problems typically require multiple solution approaches and the integration of capabilities across mathematics and physics subdisciplines. Participants in these competitions represent some of the world's strongest talent in physics logical reasoning. The problems often combine theoretical understanding with practical applications, requiring students to demonstrate both analytical and creative problem-solving skills. The international nature of these competitions ensures a diverse range of problem-solving approaches and cultural perspectives.

- **Chinese National College Entrance Examination (Gaokao) Physics Questions**
  The Gaokao physics questions represent a rigorous standardized assessment system that has been refined over decades. These questions are designed to test both fundamental understanding and advanced application of physics concepts at the high school level. The problems are carefully calibrated to discriminate between different levels of student ability while maintaining high reliability and validity. They often incorporate real-world scenarios and practical applications, making them particularly valuable for assessing applied physics knowledge.

- **Chinese Mock Examinations at Various Levels**
  Our collection includes a comprehensive range of mock examination questions from multiple administrative levels in China. This includes provincial-level mock exams, city-level assessment materials, and joint examination papers created through collaboration

among multiple high schools. These diverse sources provide a rich spectrum of problem-solving scenarios and difficulty levels. The multi-tiered nature of these mock examinations reflects different regional interpretations of educational standards while maintaining alignment with national requirements. The variety in question sources ensures exposure to different testing styles and pedagogical approaches, making this dataset particularly valuable for understanding the breadth of physics education assessment in China.

- **Indian Joint Entrance Examination (Advanced)**
  This examination represents one of India's most prestigious and challenging engineering entrance tests. The exam structure, consisting of two papers with 50-60 questions each, provides a comprehensive assessment of physics knowledge alongside mathematics and chemistry. The questions are known for their analytical depth and often require multi-step problem-solving approaches. The exam's high stakes nature and competitive environment ensure that the problems are both challenging and discriminating, making them valuable additions to our dataset.

- **Others**
  We also obtained some physics questions from non-Chinese and English sources on huggingface, such as Russian (Timur et al., 2024). This dataset consists of a diverse collection of physics problems, categorized into different domains, including 1000 problems on Kinematics, 600 problems on Electricity and Circuits, and 500 problems on Thermodynamics. All data has been extracted from open sources, ensuring a wide variety of problem types and difficulty levels.

The PhysReason benchmark is derived from publicly available physics education materials including: International Physics Olympiad problems (2008-2021), Chinese National College Entrance Examination physics questions (2010-2024), Indian Joint Entrance Examination Advanced physics problems (2010-2024), Chinese provincial and municipal mock examination questions (2015-2024). We have collected more than 20,000 physics problems. All problems were collected in accordance with fair use principles for educational and research purposes. The complete benchmark and associated code will be released under the MIT License for research use.

The dataset contains no personally identifiable information. All problems are from standardized tests and competition materials with no individual student data. This documentation ensures reproducibility and proper usage of the benchmark while protecting privacy and intellectual property rights.

B Benchmark

B.1 Collection

B.1.1 Data Acquisition

We systematically collected, curated, and processed physics problems from diverse sources to ensure comprehensive coverage of physics concepts and problem-solving scenarios. Our dataset comprises 1,254 PDF documents totaling 27,874 pages, yielding over 20,000 unique problems. This extensive collection provides a rich foundation for developing a comprehensive physics problem benchmark.

B.1.2 Data Standardization

We implemented a systematic data processing pipeline utilizing MinerU (Wang et al., 2024a) for PDF parsing. The standardization process encompasses several critical phases: initial format conversion, rigorous deduplication, and comprehensive formatting standardization. Each question underwent a rigorous quality assessment process with specific evaluation criteria:

1. Complete problem statements with well-defined variables and conditions
2. Clear and unambiguous wording
3. Accurate expressions and units
4. Consistent formatting of equations and symbols

B.1.3 Translation

To standardize the multilingual dataset comprising Chinese, English, Hindi, and Russian content, we implement a two-phase process:

Phase I: Translation

- Initial translation by translators
- Strict adherence to standardized physics terminology
- Consistent mathematical notation and expressions

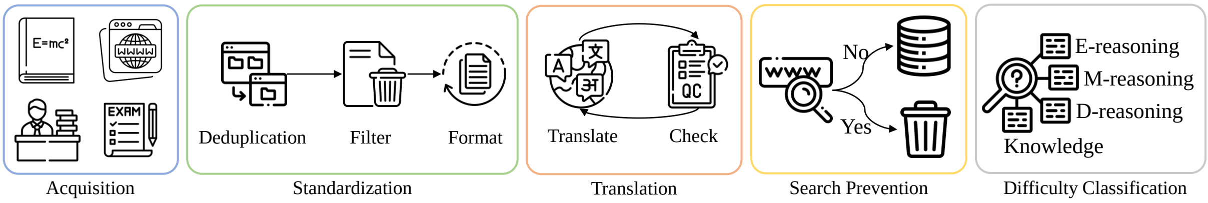


Figure 6: Illustration of the data collection pipeline.

Phase II: Verification

- Validation by Engineering Ph.D. candidates
- Verification of physics terminology accuracy
- Confirmation of semantic equivalence
- Review of mathematical expression consistency

B.1.4 Search Prevention

Following (Rein et al., 2024), we exclude problems whose answers could be found through a five-minute Google search to minimize data leakage. This step ensures that model evaluation reflects genuine physics problem-solving capabilities rather than information retrieval abilities.

B.1.5 Difficulty Classification:

Questions were systematically categorized using a multi-dimensional classification framework:

Primary Classification

- Knowledge-based questions:
  - Focus on fundamental physics concepts
  - Clear-cut application of specific theorems
  - Direct calculation or concept identification
- Reasoning-based questions:
  - Multiple-theorem integration
  - Multi-step problem-solving approaches
  - Complex analytical thinking

Difficulty Levels in Reasoning-based Questions

- Easy:
  - Total steps ≤ 5
  - Completion time: 0-5 minutes
- Medium:
  - Total steps ≤ 10
  - Completion time: 5-15 minutes
- Hard:
  - Total steps > 10
  - Completion time: 15+ minutes

B.2 Annotation

Key Elements As shown in Figure 1, our annotation framework consists of 7 key elements:

- Context:
  - Detailed physics scenario description: Describe the physics setup thoroughly, including objects, environment, and interactions. For example, specify angles, materials, initial conditions, and forces.
  - Clear specification of conditions and constraints: Explicitly list all given conditions: initial conditions (e.g., initial velocity, position), boundary conditions, and constraints (e.g., inextensible string, frictionless surface).
  - Standardized notation for physics quantities: Use consistent and standard symbols for physics quantities (e.g., v for velocity, a for acceleration, m for mass) throughout the annotation.
- Sub-question:
  - Hierarchical structure of related questions: Break down a complex problem into smaller, logically connected subquestions. These should build upon each other.
  - Clear progression of complexity: Subquestions should increase in difficulty, guiding the learner from basic concepts to more advanced analysis.
- Solution:
  - Detailed step-by-step reasoning process: Provide a comprehensive, step-by-step solution. Do not skip any crucial reasoning steps.

  - Each step contains at least one formula: Each step in the solution should include at least one relevant physics formula (theorem, law, or derived equation).
  - If the formula can be solved to a value, it should also have a value: If a step's formula yields a numerical result, provide that result.

- Step Analysis:
  - Explicit theorem application rationale: Clearly state which theorem, law, or principle is applied in each step and why it's applicable. Example: Newton's Second Law.
  - Physics quantity derivation explanation: Explain how unknown physics quantities are derived from known ones. Example: "$W = \Delta E_k$"

- Answer:
  - Numerical results with appropriate units: Provide the correct numerical value and units for numerical answers (e.g., "$v = 5m/s$").
  - Formulaic results with appropriate symbols: For formulaic answers, use previously defined standard symbols and ensure the formula's correctness (e.g., "$v = \sqrt{2gh}$").

- Difficulty:
  - Reasoning difficulty metrics: Use qualitative descriptions (e.g., "knowledge" "easy," "medium," "hard")

- Theorem:
  - Comprehensive list of applicable theorems, laws, and formulas: Provide a complete list of all the specific physics theorems, laws, and equations that are relevant to solving the problem. Examples include: 'Newton's Second Law', 'Work-Energy Theorem', 'Conservation of Momentum', 'Kinematic Equations', etc.
  - Core Concepts: Identify the fundamental physics principles and ideas that underpin the solution, even if they aren't expressed as a single equation. Examples include: 'Wave-Particle Duality'.

C Error Type Details

The following is a summary of the error types, categorized and with expanded explanations:

1. **Diagram Analysis Errors:**
   - *Description:* Errors related to the comprehension, plotting, analysis, or extraction of data from graphical representations. This encompasses any mistake made when working with diagrams, charts, or graphs.
   - *Examples:*
     - Misreading the labels or units on the axes of a graph.
     - Misinterpreting the trend of a curve (e.g., confusing a linear relationship with an exponential one).
     - Failing to identify key data points or features on the graph (e.g., maxima, minima, intercepts).
     - Incorrectly extrapolating or interpolating data from the graph.
     - Drawing an inaccurate graph based on given data.

2. **Physics Theorem Application Errors:**
   - *Description:* Errors arising from the incorrect application of physics theorems or principles, or using them in situations where they are not valid. This includes both misremembering the law itself and misapplying a correctly remembered law.
   - *Examples:*
     - Applying Newton's Laws of Motion to a non-inertial reference frame without accounting for fictitious forces.
     - Using the conservation of energy principle in a system where non-conservative forces (like friction) are doing significant work.
     - Applying a formula outside of its valid range of applicability (e.g., using a small-angle approximation when the angle is large).
     - Misunderstanding the conditions under which a particular law is valid.

3. **Physics Condition Analysis Errors:**

- *Description:* Errors related to the incorrect assessment of the physics system's boundaries, the forces acting on it, or its constituent components. This involves a misunderstanding of 'what' is happening in the system.
- *Examples:*
  - Neglecting the force of friction in a situation where it is significant.
  - Incorrectly identifying the system boundary, leading to errors in applying conservation laws.
  - Misjudging whether a system is isolated (no external forces) or not.
  - Failing to consider all relevant forces acting on an object.
  - Misidentifying the components of a system that are interacting.

4. **Physics Process Understanding Errors:**

- *Description:* Errors stemming from a misunderstanding of how a physics phenomenon develops, how states change, or the causal relationships between events. This involves a misunderstanding of 'how' things are happening.
- *Examples:*
  - Incorrectly analyzing the motion of a projectile, such as misunderstanding the independence of horizontal and vertical motion.
  - Misunderstanding the mechanisms of energy transformation (e.g., confusing heat and temperature).
  - Incorrectly predicting the direction of motion based on the forces involved.
  - Having misconceptions about the nature of a physics process (e.g., believing that a continuous force is needed to maintain constant velocity).

5. **Variable Relationship Errors:**

- *Description:* Errors caused by misunderstanding the dependencies or functional relationships between different physics quantities. This involves incorrectly relating variables.
- *Examples:*
  - Incorrectly assuming that acceleration is directly proportional to velocity.
  - Misunderstanding the relationship between force, mass, and acceleration (Newton's Second Law).
  - Confusing the relationship between potential and kinetic energy.
  - Failing to recognize an inverse relationship between two variables.

6. **Calculation Process Errors:**

- *Description:* Errors occurring during the mathematical manipulation of equations, the derivation of formulas, or the substitution of numerical values. These are purely mathematical mistakes.
- *Examples:*
  - Making algebraic errors when rearranging equations.
  - Incorrectly performing unit conversions (e.g., mixing up meters and centimeters).
  - Making arithmetic errors (e.g., simple addition or multiplication mistakes).
  - Incorrectly substituting values into a formula.
  - Errors in using a calculator.

7. **Boundary Condition Analysis Errors:**

- *Description:* Errors resulting from neglecting or mishandling special cases, limiting conditions, or the applicable ranges of variables or equations. This involves not considering the "edges" of the problem.
- *Examples:*
  - Failing to consider the behavior of a system at extremely high or low temperatures.
  - Neglecting the effects of air resistance when analyzing projectile motion at high speeds.
  - Not considering the limitations of a particular model or approximation.
  - Applying a formula outside its range of validity.
  - Ignoring initial conditions or other constraints.

D Example

We have provided a representative example for each of the four question difficulty levels—knowledge (Figure 7), easy (Figure 8), medium (Figure 9), and hard (Figure 10) to serve as a guide.

The knowledge-level problem demonstrates the fundamental application of electromagnetic principles, requiring direct use of basic physics theorems without complex problem-solving steps. This type of question focuses on testing models' understanding of core concepts and their ability to apply basic formulas.

The easy-level problem involves a straightforward mechanical system with clear physics conditions. It requires models to apply basic conservation laws and Newton's laws in a sequential manner, with each step building logically on the previous one. The solution path is direct and requires minimal manipulation.

The medium-level problem introduces multiple state changes and requires models to analyze a system under different configurations. It combines several physics principles and demands a more sophisticated understanding of how different variables interact. The solution requires models to track system changes systematically while maintaining consistency in their physics-based reasoning.

The hard-level problem presents a complex mechanical system with multiple connected components and sequential events. It requires models to analyze a series of interactions, apply multiple physics principles simultaneously, and consider various constraints throughout the problem-solving process. The solution demands both careful physics insight and mathematical rigor, testing models' ability to synthesize different concepts and handle multi-step calculations.

These examples demonstrate the progressive complexity in physics problem-solving across different difficulty levels. From knowledge-level questions testing basic concept application, to hard problems requiring integration of multiple physics principles and sophisticated analysis, each level builds upon the previous one. This hierarchical structure effectively assesses models' comprehension and problem-solving abilities, ranging from fundamental understanding to advanced physics-based reasoning and mathematical manipulation. The gradual increase in complexity helps evaluate models' mastery of both individual concepts and their ability to synthesize multiple physics principles in complex scenarios.

E Evaluation Prompt

To systematically evaluate models' mathematical reasoning capabilities, we designed a structured prompt template that follows the zero-shot Chain-of-Thought (CoT) paradigm. This template adopts a hierarchical structure comprising image information, problem context, and sequential sub-questions, requiring models to provide standardized step-by-step solutions. The prompt structure consists of the following key components:

E.1 Input Components

- **Image Caption:** For models without direct image processing capabilities, we utilize Gemini-2.0-flash to generate image descriptions as supplementary information
- **Context:** Provides the overall background and fundamental information of the problem
- **Sub-questions:** Decomposes complex problems into progressive sub-questions

E.2 Output Specifications

The template requests a structured output format with the following requirements:

- Step-by-step reasoning for each sub-question
- Continuous step numbering across sub-questions
- One formula and its solution process per step
- Mathematical formulas enclosed in LaTeX notation ($)

This design adheres to the zero-shot Chain-of-Thought paradigm, facilitating systematic thinking through explicit step division and standardized output format, which benefits both model reasoning and subsequent performance evaluation. The template's flexibility allows it to accommodate pjhysical problems of varying complexity, with adjustable numbers of sub-questions and solution steps based on specific problem requirements.

F Details of Experimental Result

We previously presented only partial model performance benchmarks on PhysReason-mini. And we provide a comprehensive performance evaluation across all models, as shown in Table 7.

Table 7: Model performance comparisons on the PhysReason-mini benchmark using answer-level evaluation across different input combinations of Questions (Q), Images (I), and Image Captions (IC). Gemini-2.0-T† and * represent Gemini-2.0-Flash-Thinking-1206 and 0121.

Model	Input	Knowledge	Easy	Medium	Hard	Avg.
Non-O-like Models						
Qwen2VL-72B	Q, I	25.40	27.00	11.4	8.5	18.07
InternVL2.5-78B	Q, I	37.90	20.60	18.14	7.97	21.15
GPT-4o	Q, I	51.12	31.95	20.75	12.54	29.09
Claude-3.5-Sonnet	Q, I	49.00	40.43	23.45	12.33	31.3
Deepseek-V3-671B	Q, IC	56.60	40.97	22.22	14.61	33.6
Gemini-2.0-Flash	Q, I	67.80	52.10	40.00	23.19	46.52
Gemini-2.0-Pro	Q, I	69.32	53.67	44.98	26.24	48.55
O-like Models						
o1-mini	Q, IC	54.80	30.33	15.41	7.92	27.11
QvQ-72B	Q, I	51.17	37.10	29.83	22.13	35.06
QwQ-32B	Q, IC	64.4	50.07	38.88	27.45	45.20
Gemini-2.0-T†	Q, I	71.47	49.97	36.83	22.97	45.42
GLM-Zero	Q, IC	72.70	50.17	43.42	24.70	47.75
o1	Q, I	72.47	53.37	49.31	25.32	50.12
o3-mini-high	Q, IC	71.10	63.20	47.02	31.93	53.31
Gemini-2.0-T*	Q, I	76.33	56.87	51.85	32.61	54.42
Deepseek-R1	Q, IC	85.17	60.77	47.24	33.23	56.60

G Details of Scientific Artifacts

Our PhysReason benchmark dataset integrates problems from multiple sources: International Physics Olympiad (2008-2021), Chinese National College Entrance Examination (2010-2024), Indian Joint Entrance Examination Advanced (2010-2024), Chinese provincial and municipal mock examination questions (2015-2024), and additional physics problems from Russian sources, totaling over 20,000 unique physics problems from 1,254 PDF documents across 27,874 pages. The dataset has been carefully curated to ensure comprehensive coverage while respecting intellectual property rights - all problems are utilized under the CC BY-NC-SA and MIT licenses, and all materials were collected in accordance with fair use principles for educational and research purposes. We maintain strict privacy standards with no personally identifiable information included, as all problems are sourced from standardized tests and competition materials. The complete benchmark and associated code are made available for research use, requiring users to comply with both the MIT License terms for our implementation and the respective original licenses (CC BY-NC-SA) for the educational materials, thereby ensuring proper attribution and usage rights while promoting academic accessibility.

H Details of Computational Experiment

Our computational experiments were conducted across multiple Large Language Models (LLMs), Vision Language Models (VLMs), and other specialized models. The infrastructure primarily consisted of NVIDIA A800 GPUs for running open-source models. For model specifications, we evaluated seventeen models in total, including Qwen2-VL-72B (72 billion parameters), QwQ-32B (32 billion parameters), QvQ-72B (72 billion parameters), InternVL2.5-78B (78 billion parameters), and various other commercial models like GPT-4, Claude-3.5-Sonnet, and Gemini series. All experiments were conducted under a zero-shot Chain-of-Thought (CoT) setting to encourage step-by-step reasoning. For the experimental setup, we utilized specific prompts (detailed in supplementary materials) to maintain consistency across all evaluations. The models processed the complete PhysReason benchmark dataset, with the exception of O1 due to API limitations. For performance evaluation, we employed both PSAS-A and PSAS-S frameworks, with Deepseek-V3 ultimately selected as the scoring model based on efficiency and performance considerations. Regarding implementation details, models that couldn't process visual inputs were supplemented with image captions generated by Gemini-2.0-Flash. For reproducibility purposes, all prompt templates are provided in the supplementary materials. Due to the computational cost of the PSAS-S framework, some experiments were conducted using only the PSAS-A framework to maintain efficiency.

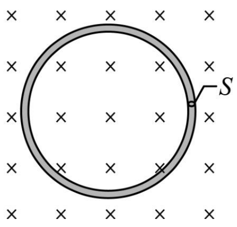
I Details of human annotators

For data annotation and evaluation, we engaged four graduate students (including both PhD and Master's students) from engineering disciplines who are also co-authors of this paper. All annotators possessed strong backgrounds in both high school and undergraduate physics, making them well-qualified for this task. Since the annotators were co-authors actively involved in the research, no formal recruitment process or compensation was required, and they were fully aware of how the data would be used in the study. The annotation process focused solely on physics content evaluation and did not involve collecting any personal identifying information or expose annotators to any risks. As this research involved co-authors analyzing academic content rather than external human subjects, it was determined to be exempt from formal ethics review board approval. The annotation work was conducted as part of regular academic research activities within our institution. No protected or sensitive demographic information was collected or used in this research.

J Details of Ai Assistants In Research Or Writing

We used Claude-3.5-Sonnet, o1, o3-mini-high, and Deepseek-R1 to help us write code and polish the paper.

Diagram:



Context:
A student investigates the thermal effects of metal objects in a changing magnetic field, using a metal ring as the subject. The ring can be considered as a single-turn circular coil with a circumference of L, a cross-sectional area of S, and a resistivity of ρ. It is placed in a uniform magnetic field with the magnetic induction direction perpendicular to the plane of the ring. The magnitude of the magnetic induction increases uniformly from 0 to B_{0} within a time interval of Δt.",

Sub-questions:
(1) What is the induced electromotive force in the ring?
(2) What is the current in the ring?
(3) What is the Joule heating power of the current in the ring?

Solution:
Sub-question-1:
step_1: Let the radius of the ring be r, then we have $L=2\pi r$.
step_2: The magnitude of the magnetic induction increases uniformly from 0 to B_{0} within a time interval Δt, and the induced electromotive force is given by $E=\frac{B_{0}}{\Delta t} \cdot \pi r^{2}$, which leads to $E={\frac{B_{0}L^{2}}{4 \pi \Delta t}}$.
Sub-question-2:
step_3: The resistance of the ring is $R=\frac{\rho L}{S}$.
step_4: The induced current in the ring is $I=\frac{E}{R}=\frac{B_{0}L S}{4 \pi \rho \Delta t}$.
Sub-question-3:
step_5: The thermal power of the current in the ring is $P=I^{2}R={\frac{B_{0}^{2}L^{3}S}{16 \pi ^{2} \rho (\Delta t)^{2}}}$.

Answer: (1) $E={\frac{B_{0}L^{2}}{4\pi\Delta t}}$, (2) $I=\frac{B_{0}L S}{4\pi\rho\Delta t}$, (3) $P=\frac{ B_{0}^{2}L^{3}S}{16\pi^{2}\rho(\Delta t)^{2}}$

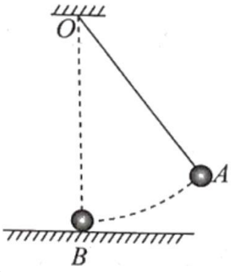
Theorem: Faraday's Law of Induction, Ohm's Law, Joule's Law

Difficulty: Knowledge

Step Analyze:
step_1:
(1) Theorem: N/A
(2) Target: Radius of the ring,
(3) Formula: $L=2\\pi r$,
(4) Value: $\\frac{L}{2\\pi}$,
step_2:
(1) Theorem : Faraday's Law of Induction
(2) Target : Induced electromotive force in the ring",
(3) Formula: $E=\frac{B_{0}}{\Delta t} \cdot \pi r^{2}$
(4) Value: $E={\frac{B_{0}L^{2}}{4\pi\Delta t}}$
step_3:
(1) Theorem : N/A
(2) Target : Resistance of the ring,
(3) Formula: R = \rho \frac{L}{S},
(4) Value: \rho \frac{L}{S},
step_4:
(1) Theorem : Ohm's Law
(2) Target : Induced current in the ring
(3) Formula: $I = \frac{E}{R}$
(4) Value: $\frac{B_{0}L S}{4\pi\rho\Delta t}$
step_5:
(1) Theorem : Joule's Law
(2) Target: Joule heating power of the current in the ring
(3) Formula: $P=I^{2}R$
(4) Value:${\frac{B_{0}^{2}L^{3}S}{16\pi^{2}\rho(\Delta t)^{2}}}$

Figure 7: A knowledge example in our benchmark.

Diagram:



Context:
A small ball A of mass m is suspended from point O by an inextensible light string. On a smooth table directly below point O, there is an identical stationary small ball B, with a distance from O equal to the length of the string, L. Ball A is now pulled to a certain height and released from rest. A collides head-on with B at a speed v in the horizontal direction and sticks together with B. The acceleration due to gravity is g.

Sub-questions:
(1) What is the height H of point A above the table when released?
(2) What is the magnitude of the tension force F in the string just before collision?
(3) How much mechanical energy ΔE is lost during the collision?

Solution:
Sub-question-1:
step_1: According to the law of conservation of mechanical energy, $m g H=\frac{1}{2}mv^{2}$, solving for H yields $H=\frac{v^{2}}{2g}$.
Sub-question-2:
step_2: From Newton's second law of motion, $F-mg=m\frac{v^{2}}{L}$, we obtain $F-m g =m\frac{v^2}{L}$.
Sub-question-3:
step_3: During the collision process, applying the law of conservation of momentum, $m v=2m v^{\prime}$, and solving for $v^{\prime}$ gives $v^{\prime}=\frac{v}{2}$.
step_4: The mechanical energy lost during the collision, ΔE, is calculated as $\Delta E ={\frac{1}{2}}m{v}^{2} - {\frac{1}{2}}\cdot 2m{v}^{\prime}^{2} = {\frac{1}{4}}m{v}^{2}$.

Answer: (1) $H=\frac{v^{2}}{2g}$, (2) $F=mg+m\frac{v^{2}}{L}$, (3) $\Delta E=\frac{1}{4}mv^{2}$

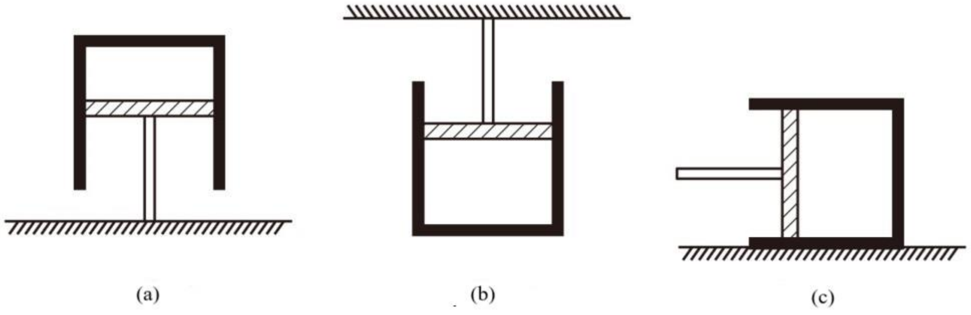
Theorem: Conservation of Energy, Second Newton's Law, Conservation of momentum, Definition of mechanical energy loss

Difficulty: easy

Step Analyze:
step_1:
(1) Theorem: Conservation of Energy
(2) Target: Height of A above the table when released
(3) Formula: $mgH = \frac{1}{2}\times m \times v^2$
(4) Value: \frac{ v^2}{2g}
step_2:
(1) Theorem: Newton's second law of motion
(2) Target: Tension in the string just before collision
(3) Formula: $F - mg = m \times \frac{v^2}{L}$
(4) Value: mg + m \times \frac{v^2}{L}
step_3
(1) Theorem: Law of conservation of momentum
(2) Target: Velocity of A and B after collision
(3) Formula: m \times v = 2 \times m \times v^{\prime}
(4) Value: \frac{v}{2},
step_4:
(1) Theorem: Definition of mechanical energy loss
(2) Target: Mechanical energy loss during the collision
(3) Formula: $\Delta E = \frac{1}{2} \times m \times v^2 - 1/2 \times 2m \times {v^{\prime}}^2$
(4) Value: \frac{1}{4}mv^2

Figure 8: An easy example in our benchmark.

Diagram:



Context:
A thin-walled, thermally conductive cylindrical container with mass m and smooth inner walls encloses a certain amount of ideal gas with a piston of cross-sectional area S. In all the following processes, the cylinder does not leak and remains in contact with the piston. When the cylinder is placed vertically inverted as shown in figure (a), the gas volume inside the cylinder is V_{1}, and the temperature is T_{1}. Given the magnitude of gravitational acceleration as g, and the atmospheric pressure as p_{0}.

Sub-questions:
(1) When the cylinder is suspended vertically as shown in figure (b) and the gas temperature remains at T_1, what is the gas volume V_2 inside the cylinder?
(2) In figure (c), after the horizontally placed cylinder reaches equilibrium and is slowly heated, what is the temperature of the gas when its volume becomes V_3?

Solution:
Sub-question-1:
step_1: In the state of Figure (a), performing force analysis on the cylinder, the pressure of the enclosed gas is $p_{1}=p_{0}+\frac{m g}{S}$
step_2: When the cylinder is suspended as shown in Figure (b), performing force analysis on the cylinder, the pressure of the enclosed gas is $p_{2}=p_{0}-{\frac{m g}{S}}$
step_3: Applying Boyle's law to the enclosed gas, we have $p_{1}V_{1}=p_{2}V_{2}$
step_4: Combining step_1, step_2, and step_3, we obtain $V_{2}=\frac{p_{0}S+m g}{p_{0}S-m g}V_{1}$
Sub-question-2:
step_5: When the cylinder is placed horizontally as shown in Figure (c), the pressure of the enclosed gas is $p_{3}=p_{0}$
step_6: From the ideal gas law, we have $\frac{p_{1}V_{1}}{T_{1}}=\frac{p_{3}V_{3}}{T_{3}}$
step_7: Combining step_1, step_5, and step_6, we obtain $T_{3}=\frac{p_{0}S V_{3}T_{1}}{(p_{0}S+m g)V_{1}}$

Answer: (1) $\frac{p_{0}S+m g}{p_{0}S-m g}V_{1}$, (2) $\frac{p_{0}S V_{3}T_{1}}{(p_{0}S+m g)V_{1}}$

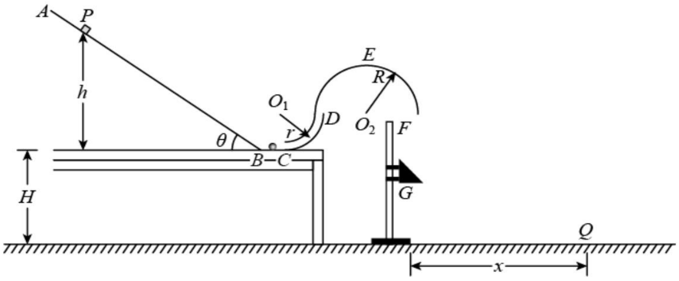
Theorem: Boyle's Law, Ideal gas law, Force equilibrium

Difficulty: medium

Step Analyze: Due to page space limitations, not displayed

Figure 9: A medium example in our benchmark.

Diagram:



Context:
A horizontal platform of height $H=0.4m$ is placed on a level ground, on which a rough straight track AB with an inclination angle of $\theta=37^{\\circ}$, a horizontal smooth straight track BC, a quarter-circular smooth thin circular pipe CD, and a semi-circular smooth track DEF are vertically placed, and they are smoothly connected. The radius of the pipe CD is $r=0.1m$ with its center at point O_{1}, and the radius of the track DEF is $R=0.2m$ with its center at point O_{2}. Points O_{1}, D, O_{2}, and F are all on the same horizontal line. A small slider starts from rest at point P on the track AB, which is at a height h above the platform, and slides down. It undergoes an elastic collision with a small ball of equal mass at rest on the track BC. After the collision, the small ball passes through the pipe CD and the track DEF, moving vertically downwards from point F and collides with a triangular prism G fixed on a straight rod directly below. After the collision, the ball's velocity direction is horizontal to the right, and its magnitude is the same as before the collision. Finally, it lands at point Q on the ground. The coefficient of kinetic friction between the slider and the track AB is $\mu=\frac{1}{12}$, $\\sin37^{\circ}=0.6$, and $\cos37^{\circ}=0.8$.

Sub-questions:
(1) When the slider falls from an initial height of $h=0.9m$, what is its velocity v_0 when it reaches point B?
(2) What is the minimum height h_{min} needed for the small ball to complete the entire motion?
(3) If the small ball just barely passes the highest point E and the triangular prism G can be adjusted vertically, what is the maximum horizontal distance x_{max} between the landing point Q and point F?

Solution:
Sub-question-1:
step_1: The motion of the small slider on the AB track is governed by the energy equation $mgh - \mu m g\cos\theta \cdot \frac{h}{\sin\theta}=\frac{1}{2}mv_{0}^{2}$. Substituting the given values, the initial velocity is calculated as $v_{0}=\frac{4}{3}sqrt{gh}=4\mathrm{m/s}$
Sub-question-2:
step_2: The small ball moves along the CDEF track. At the highest point, the centripetal force is equal to the gravitational force, yielding $m g=m\frac{v_{(E, min)}^{2}}{R}$
step_3: Applying the conservation of mechanical energy from point C to point E, we obtain $\frac{1}{2}mv_{(E, min)}^{2}+mg(R+r)=\frac{1}{2}mv_{(B, min)}^{2}$
step_4: Based on step_2 and step_3, the minimum velocity at point E is $v_{(E, min)}=\sqrt{2}\mathrm{m/s}$ and the minimum velocity at point B is $v_{(B, min)}=2\sqrt{2}\mathrm{m/s}$
step_5: After the collision between the small slider and the small ball, the momentum is conserved, which gives $mv_{(A, min)}=mv_{(A, min)}^{\prime}+mv_{(B, min)}$.
step_6: After the collision between the small slider and the small ball, the mechanical energy is conserved, which gives $\frac{1}{2}mv_{(A, min)}^{2} = \frac{1}{2}mv_{A}^{\prime}^{2} + \frac{1}{2} mv_{(B, min)}^{2}$.
step_7: Based on step_5 and step_6, we obtain $v_{A}^{\prime}=0$ and $v_{(B, min)}=v_{(A, min)}$.
step_8: The motion of the small slider on the AB track is determined by $mgh - \mu mg \cos \theta \cdot \frac{h}{\sin\theta} =\frac{1}{2}mv_{0}^{2}$, and we have $v_{(A, min)}=\frac{4}{3}\sqrt{g h_{min}}$.
step_9: Thus, combine step_5, step_6, step_7 and step_8, we can get $h_{min}=0.45\mathrm{m}$.
Sub-question-3:
step_10: Let the distance from point F to point G be denoted as y. For the motion of the small ball from point E to point G, the work-energy theorem gives $\frac{1}{2}mv_{G}^{2}=\frac{1}{2}mv_{E{min}}^{2}+m g(R+y)$.
step_11: From the projectile motion, we can derive $x=v_{G}t$.
step_12: And we can also get $H+r-y=\frac{1}{2}g t^{2}$.
step_13: Combining step_9 and step_10, we have $x=2\\sqrt{(0.5-y)(0.3+y)}$. The value of x reaches a maximum when $0.5-y=0.3+y$.
step_14: Therefore, combining step_11, step_12, step_13, the maximum value of x is $x_{max}=0.8\mathrm{m}$.

Answer: (1) $4m/s$, (2) $h_{min} = 0.45m$, (3) $ x_{max}=0.8m$

Difficulty: difficult

Theorem: Work-energy theorem, Newton's second law, Conservation of momentum, Conservation of mechanical energy

Step Analyze: Due to page space limitations, not displayed

Figure 10: A hard example in our benchmark.