PHYSICSARENA: The First Multimodal Physics Reasoning Benchmark Exploring Variable, Process, and Solution Dimensions

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated remarkable capabilities in diverse reasoning tasks, yet their application to complex physics reasoning remains underexplored. Physics reasoning presents unique challenges, requiring grounding in physical conditions and the interpretation of multimodal information. Current physics benchmarks are limited, often focusing on *text-only inputs* or *solely on problem-solving*, thereby overlooking the critical intermediate steps of variable identification and process formulation. To address these limitations, we introduce PHYSICSARENA, the first multimodal physics reasoning benchmark designed to holistically evaluate MLLMs across three critical dimensions: *variable identification*, *physical process formulation*, and *solution derivation*. PhysicsArena aims to provide a comprehensive platform for assessing and advancing the multimodal physics reasoning abilities of MLLMs.

1 Introduction

Multimodal Large Language Models (MLLMs) have recently demonstrated remarkable capabilities across a diverse range of domains (Caffagni et al., 2024; Fei et al., 2024; Yan et al., 2024c,b). Their proficiency in processing and integrating information from various modalities has unlocked significant potential (Fu et al., 2024; Huo et al., 2024). Notably, the reasoning abilities inherent in the underlying LLMs have fueled advancements in multimodal reasoning tasks. This synergy is particularly beneficial in complex, real-world scenarios such as education, where understanding and reasoning about multimodal information are paramount. Areas like mathematical problem-solving and code generation have already seen substantial progress, showcasing the power of MLLMs in tackling structured reasoning challenges (Yan et al., 2024a; Yun et al., 2024; Wang et al., 2024a; Lin et al., 2025).

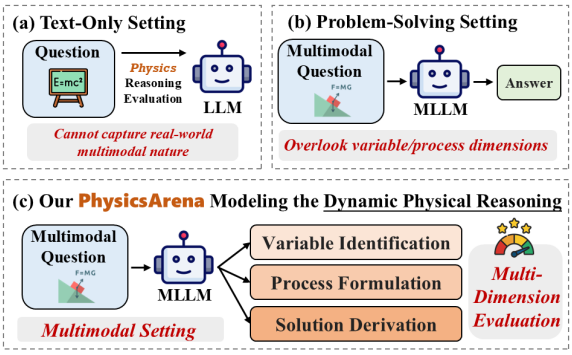


Figure 1: Comparison between previous physics reasoning settings and our proposed PHYSICSARENA.

Despite these advancements, the *domain of physics reasoning remains relatively underexplored within the MLLM research landscape*. Physics presents a unique and arguably more intricate reasoning setting compared to mathematics or coding. Effective physics reasoning necessitates not only logical deduction but also a deep understanding **grounded in real-world physical laws and theorems**. Furthermore, the reasoning process is often tightly **constrained by objective physical conditions** depicted visually or described textually. This inherent complexity, involving the interplay between abstract principles and concrete, often multimodal, scenarios, necessitates a dedicated benchmark capable of rigorously evaluating the physics reasoning capabilities of modern MLLMs (Yan et al., 2025a; Ferrag et al., 2025).

Current benchmarks designed for physics reasoning suffer from significant limitations as follows. ❶ Many existing efforts (Qiu et al., 2025; Xu et al., 2025) primarily focus on text-only settings, *failing to capture the crucial interplay with multimodal information that characterizes real-world physics problems* (e.g., interpreting diagrams, graphs, or experimental setups), as shown in Figure 1 (a). ❷ Other benchmarks (Feng et al., 2025; Zhang et al., 2025), even if multimodal, tend to *concentrate solely on the problem-solving aspect* – predicting

the final solution or answer, as shown in Figure 1 (b). This overlooks the critical intermediate steps inherent in physics reasoning: identifying relevant variables from the problem context and formulating the correct physical process or sequence of principles required to reach the solution. A comprehensive evaluation of physics reasoning capabilities, therefore, requires modeling the dynamic reasoning process from its inception, encompassing these vital variable and process stages.

To bridge this gap, we introduce **PHYSICSARENA, the first benchmark specifically designed to comprehensively evaluate multimodal physics reasoning** across three crucial dimensions: *Variable Identification*, *Process Formulation*, and *Solution Derivation*. As illustrated in Figure 1 (c), PHYSICSARENA provides a structured environment with problems presented multimodally, demanding that models demonstrate understanding throughout the entire reasoning pipeline, not just at the final output stage. By dissecting the reasoning task into these three interconnected dimensions, our benchmark offers a more granular and insightful assessment of MLLM capabilities in this challenging domain. We have rigorously evaluated a suite of representative, state-of-the-art MLLMs using PHYSICSARENA.

Our contributions can be summarized as follows:

- We introduce PHYSICSARENA, the first multimodal physics reasoning benchmark that explicitly models the dynamic reasoning process. It comprises over 5,000 high-quality instances.
- PHYSICSARENA provides a holistic evaluation framework by incorporating assessments across the Variable, Process, and Solution dimensions. This multi-faceted approach fully addresses the complexity inherent in the physical setting.
- We conduct extensive experiments on representative state-of-the-art MLLMs using PHYSICSARENA. Our results provide valuable insights into capabilities, revealing a significant gap that still exists towards AGI-level intelligence.

2 Related Works

2.1 Physics Reasoning Benchmarks

As the community's focus on scientific reasoning increases (Luo et al., 2025; Yan et al., 2025b; Yan and Lee, 2024), physics reasoning also requires high-quality benchmarks for evaluation. As indicated in Table 1, early physics reasoning data were all subsets of general scientific reasoning benchmarks. Early science-wide suites such as E-EVAL (Hou et al., 2024) for Chinese K-12 education, MMLU-Pro (Wang et al., 2024b) for college-level knowledge, and the multimodal ScienceQA dataset (Lu et al., 2022) establish broad coverage with text-only or image-augmented multiple-choice questions across diverse subjects that include physics. Subsequent resources raise disciplinary depth: GPQA (Rein et al., 2024) introduces graduate-level STEM questions designed to be *Google-proof*; JEEBench (Arora et al., 2023) curates IIT·JEE-Advanced problems combining MC and open-ended formats; and college-focused sets such as SciBench (Wang et al., 2023), SciEval (Sun et al., 2024), and the bilingual multimodal OlympiadBench (He et al., 2024) adopt numerical or free-response answers and often supply diagram contexts. In the past year, reasoning benchmarks specifically dedicated to physics have begun to emerge. PhysReason (Zhang et al., 2025) provides 1,200 problems with step-level assessment, PHYBench (Qiu et al., 2025) introduces an expression-distance metric over 500 real-world scenarios, and UGPhysics (Xu et al., 2025) couples 5,520 undergraduate problems with a rule-based judgment pipeline. Together these benchmarks trace a coherent evolution from general science to domain-focused physics, from fixed-choice to open-ended solutions, and from text to richly multimodal settings (Chen et al., 2025a; Li et al., 2025).

2.2 Multimodal Large Language Models

Research on MLLMs has progressed from add-on visual interfaces to tightly unified vision–language architectures. Early adapters such as Visual-GPT (Wu et al., 2023), which grafts a self-reviving visual encoder onto GPT2 for data-efficient captioning, and GPT-4o (OpenAI et al., 2023), which simply enables image input for a general-purpose LLM, showed that pre-trained text decoders can address visual tasks. Later work pursues deeper fusion: Flamingo (Alayrac et al., 2022) bridges frozen vision and language backbones with cross-attention, while BLIP-2 (Li et al., 2023) links off-the-shelf encoders through a lightweight querying transformer. Leveraging CLIP's contrastive alignment of image–text embeddings (Radford et al., 2021), LLaVA (Liu et al., 2023a) feeds CLIP visual tokens directly into a chat-oriented LLM for unified multimodal reasoning. Scaling

Benchmarks	Size	Img.#	Know. Level	Qns. Type	Task Dimension		
					Variable	Process	Solution
E-EVAL (Hou et al., 2024)	342	0	K12	MC	✗	✗	–
MMLU-Pro (Wang et al., 2024b)	1299	0	COL	MC	✗	✗	–
GPQA (Rein et al., 2024)	227	0	Ph.D	OE	✗	✗	✓
JEEBench (Arora et al., 2023)	123	0	CEE	OE/MC	✗	✗	✓
ScienceQA (Lu et al., 2022)	1923	1328	K12	MC	✗	✗	–
SciBench (Wang et al., 2023)	291	64	COL	OE	✗	✗	✓
MMMU (Yue et al., 2024)	443	443	COL	OE/MC	✗	✗	✓
OlympiadBench (He et al., 2024)	2334	1958	CEE/COMP	OE	✗	✗	✓
SciEval (Sun et al., 2024)	1657	0	–	OE/MC	✗	✗	✓
EMMA (Hao et al., 2025)	156	156	CEE	MC	✗	✗	–
PhyReason (Zhang et al., 2025)	1200	972	CEE/COMP	OE	✗	✗	✓
PHYBench (Qiu et al., 2025)	500	0	COMP/COL	OE	✗	✗	✓
PHYSICS (Feng et al., 2025)	1297	298	COL	OE	✗	✗	✓
UGPhysics (Xu et al., 2025)	5520	0	UG	OE/MC	✗	✗	✓
PHYSICSARENA (Ours)	5103	5103	CEE	OE	✓	✓	✓

Table 1: Comparisons between physics reasoning benchmark (covering the physics-related data included in scientific reasoning benchmarks) vs our proposed PHYSICSARENA dataset. **Img.#:** Count of problems with image; **Knowledge Level:** *K12*: Elementary to High School; *CEE*: College Entrance Examination; *COMP*: Competition; *COL*: College; *UG*: Undergraduate; *Ph.D*: Doctor of Philosophy. **Question Type:** *OE*: Open-ended; *MC*: Multiple-choice.

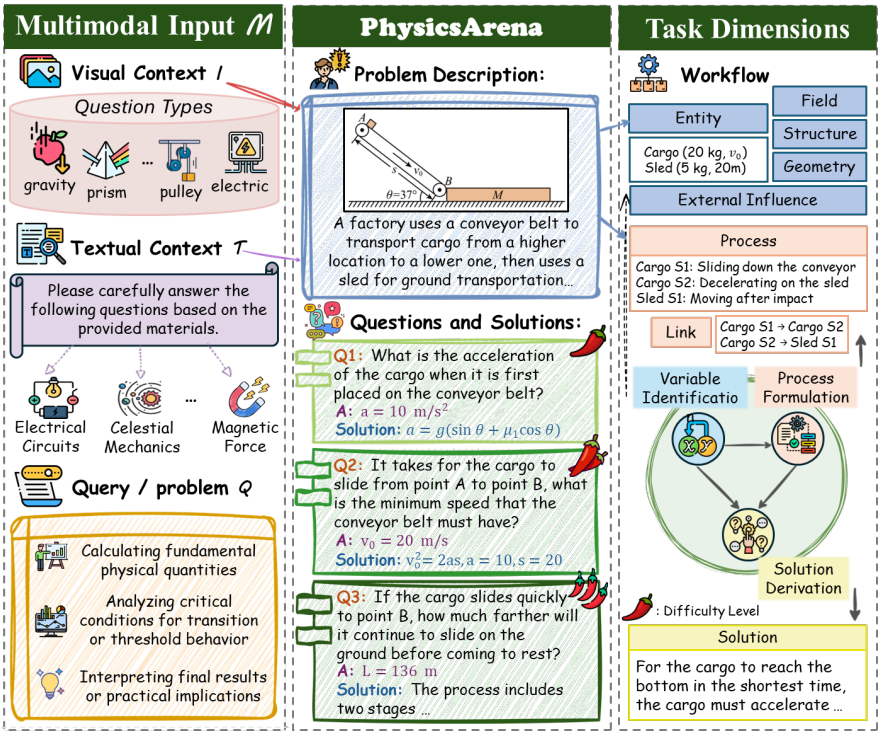


Figure 2: The illustration of a representative example from our proposed PHYSICSARENA dataset.

this paradigm, Qwen-VL (Bai et al., 2023) and InternVL (Chen et al., 2024b) co-train large vision–language encoders and attain state-of-the-art results across captioning, VQA and grounding. DeepSeek-VL (Lu et al., 2024) further introduces a hybrid multiscale vision backbone that preserves linguistic fluency while processing high-resolution images. Collectively, these works chart a clear trend toward instruction-tuned MLLMs that operate in a shared semantic space across modalities. The PHYSICSARENA dataset we propose serves as a comprehensive evaluation base for the latest representative MLLMs.

3 Our PHYSICSARENA Benchmark

3.1 Task Formulation

The core objective of PHYSICSARENA is to provide a comprehensive framework for evaluating the mul-

timodal physics reasoning capabilities of MLLMs. As shown in Figure 2, each problem instance in PHYSICSARENA is represented by a multimodal input, denoted as $\boldsymbol{M}$. This input $\boldsymbol{M}$ comprises three key components: a Visual Context $\boldsymbol{I}$ (e.g., diagrams, experimental setups), a Textual Context $\boldsymbol{T}$ (e.g., problem descriptions, conditions), and a specific Query $\boldsymbol{Q}$ related to the physics scenario, such that $\boldsymbol{M} = (\boldsymbol{I}, \boldsymbol{T}, \boldsymbol{Q})$.

Given a multimodal input $\boldsymbol{M}_i$ for a problem instance, an MLLM is tasked to generate a structured output that demonstrates its understanding across three key dimensions: Variable Identification, Process Formulation, and Solution Derivation. The model's overall output for an instance $\boldsymbol{M}_i$ can be represented as $\boldsymbol{O}_i = (\boldsymbol{O}_{V,i}, \boldsymbol{O}_{P,i}, \boldsymbol{O}_{S,i})$, corresponding to the outputs for these three dimensions. The ground truth annotations for the same instance are denoted as $\boldsymbol{G}_i = (\boldsymbol{G}_{V,i}, \boldsymbol{G}_{P,i}, \boldsymbol{G}_{S,i})$. The evaluation of the model's output $\boldsymbol{O}_i$ against the ground truth $\boldsymbol{G}_i$ is performed by a judge function $J(\cdot, \cdot)$, implemented using GPT-4o.

For **Variable Identification**, the model is required to identify $N_V = 6$ predefined categories of physical variables from the input $\boldsymbol{M}_i$. The model's output for this subtask is $\boldsymbol{O}_{V,i} = \{o_{v,1}, o_{v,2}, \ldots, o_{v,N_V}\}$, where each $o_{v,j}$ corresponds to one of the following components: (1)**Entity**, (2)**Geometry**, (3)**Field**, (4)**Structure**, (5)**Connection**, and (6)**External Influence**. The ground truth is $\boldsymbol{G}_{V,i} = \{g_{v,1}, g_{v,2}, \ldots, g_{v,N_V}\}$. Each identified component $o_{v,j}$ is compared with its corresponding ground truth $g_{v,j}$ by the judge, which assigns a boolean score $s_{v,j} = J(o_{v,j}, g_{v,j}) \in \{\text{TRUE}, \text{FALSE}\}$.

For **Process Formulation**, the model must describe the physical process by formulating $N_P = 5$ types of descriptors. The model's output for this subtask is $\boldsymbol{O}_{P,i} = \{o_{p,1}, o_{p,2}, \ldots, o_{p,N_P}\}$, where each $o_{p,k}$ corresponds to one of the following descriptors: (1)**Entity State**, (2)**Process Detail**, (3)**Force & Energy**, (4)**State Change**, and (5)**Process Relation**. The ground truth is $\boldsymbol{G}_{P,i} = \{g_{p,1}, g_{p,2}, \ldots, g_{p,N_P}\}$. Each formulated descriptor $o_{p,k}$ is compared against its ground truth $g_{p,k}$ by the judge, which assigns a boolean consistency score $s_{p,k} = J(o_{p,k}, g_{p,k}) \in \{\text{TRUE}, \text{FALSE}\}$.

For **Solution Derivation**, the model is required to generate a detailed, step-by-step reasoning chain $\boldsymbol{O}_{S,i}$ that leads to the final answer for the query $\boldsymbol{Q}$ in the input $\boldsymbol{M}_i$. The ground truth is a reference step-by-step solution $\boldsymbol{G}_{S,i}$. The model's generated solution $\boldsymbol{O}_{S,i}$ is compared with the ground truth solution $\boldsymbol{G}_{S,i}$ for logical coherence and correctness of each step by the judge, which assigns an overall boolean score $s_{S,i} = J(\boldsymbol{O}_{S,i}, \boldsymbol{G}_{S,i}) \in \{\text{TRUE}, \text{FALSE}\}$ based on exact agreement of the entire reasoning chain.

The performance on each dimension is quantified using accuracy metrics. The accuracy for Variable Identification, Accuracy_V, is calculated as the proportion of correctly identified components:

$$\text{Accuracy}_V = \frac{1}{N_V} \sum_{j=1}^{N_V} \mathbb{I}(s_{v,j} = \text{TRUE}), \quad (1)$$

where $\mathbb{I}(\cdot)$ is the indicator function. Similarly, the accuracy for Process Formulation, Accuracy_P, is

$$\text{Accuracy}_P = \frac{1}{N_P} \sum_{k=1}^{N_P} \mathbb{I}(s_{p,k} = \text{TRUE}). \quad (2)$$

The accuracy for Solution Derivation, Accuracy_S, is directly given by

$$\text{Accuracy}_S = \mathbb{I}(s_{S,i} = \text{TRUE}). \quad (3)$$

This multi-dimensional task formulation allows PHYSICSARENA to comprehensively assess an MLLM's ability to not only predict a final answer but also to understand the underlying physical variables and processes involved in addressing the query $\boldsymbol{Q}$ based on the multimodal context $(\boldsymbol{I}, \boldsymbol{T})$.

3.2 Data Preparation & Enhancement

The construction of the PHYSICSARENA benchmark is a meticulous multi-stage process, designed to ensure the dataset's quality and utility for multimodal physics reasoning. This comprehensive endeavor encompasses four primary stages: initial data collection, rigorous preprocessing, AI-assisted expert annotation, and a final meticulous sampling review (details in Appendix A).

Data Collection We systematically gathered diverse high-school physics problems, employing custom Python spiders to harvest textual components (stems, options, solutions, answers) and associated visual materials (diagrams, formula images). This collection supports the benchmark's multimodal nature, encompassing various question types like those involving gravity, prisms, *etc*.

Preprocessing Raw data underwent extensive preprocessing, including HTML cleaning with regular expressions and GPT-4o, and OCR for formula

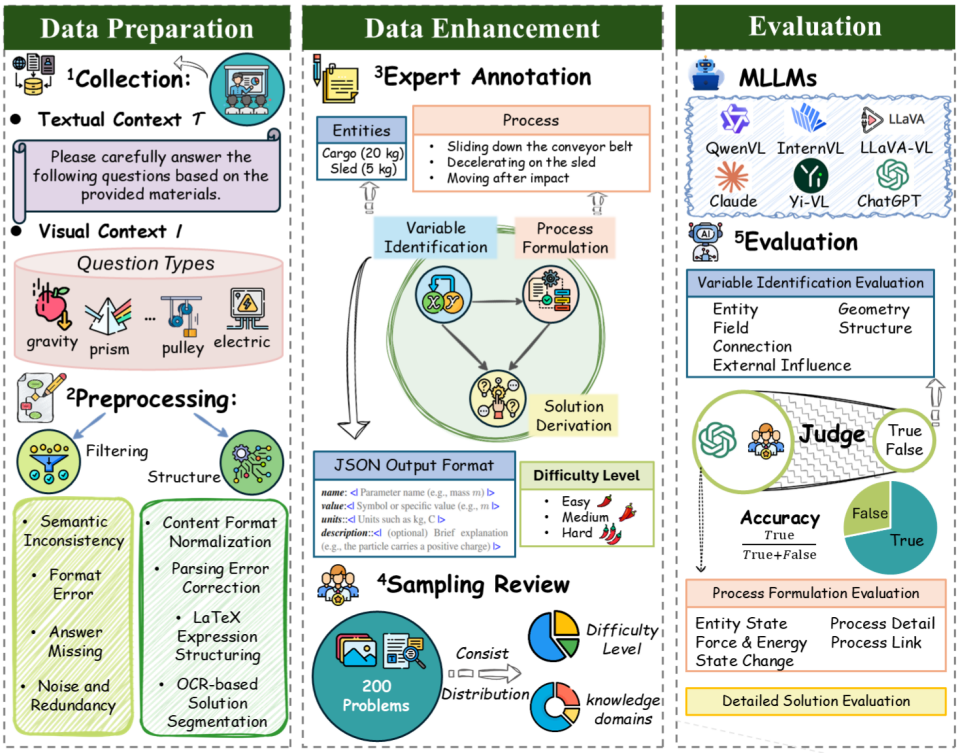


Figure 3: Roadmap of PHYSICSARENA dataset preparation, enhancement, and evaluation.

images to reconstruct LaTeX expressions. This rigorous filtering and structuring addressed inconsistencies and errors, excluded declarative knowledge items, and removed low-quality images, ensuring data integrity and a focus on procedural reasoning.

Expert Annotation Cleaned and structured data was enriched through expert annotation, leveraging GPT-4o with designed prompts (see Appendix B) to automatically generate detailed JSON annotations for each problem. These annotations specified relevant variables (entities, properties, values/units) and the formulation of physical processes, and assigned a difficulty level (Easy, Medium, Hard) to each problem.

Sampling Review Finally, a stratified subset of 200 items, reflecting original distributions of knowledge domains and difficulty, was selected for thorough manual review. Human experts meticulously examined these items to verify the accuracy of annotations, particularly for variable identification and process formulation, and to ensure the quality and reliability of PHYSICSARENA.

3.3 Dataset Details

The PHYSICSARENA dataset, as summarized in Table 2, encompasses a total of 5,103 multimodal physics problems. These problems are distributed across three distinct difficulty levels: Easy (40.7%, 2,077 items), Medium (36.2%, 1,847 items), and Hard (23.1%, 1,179 items), ensuring a comprehensive range of challenges. The dataset further exhibits broad topical coverage, with significant representation from areas such as Magnetic Fields (21.2%), Electromagnetic Induction (20.8%), and Newton's Laws of Motion (19.8%), alongside a diverse array of other fundamental physics concepts. The sample problems are presented in Appendix C.

4 Experiments and Analysis

4.1 Evaluation Protocols

We employ **GPT-4o** as the automatic judge for PHYSICSARENA. The evaluation is divided into three complementary subtasks that together assess the *structural* and *procedural* quality of a model's physical reasoning. See details of evaluation protocols and prompts in Appendix D and E.

4.2 Experimental Setup

We conduct a comprehensive evaluation on a diverse set of state-of-the-art MLLMs. Our open-source set includes **InternVL 2.5** series (Chen et al., 2025b), **Qwen 2.5-VL** series (Bai et al., 2025b), **LLaVA v1.6** (Liu et al., 2023b), and **Yi-VL-6B** (01.AI Team, 2024). For closed source, we evaluate three leading models: **GPT-4o** (OpenAI et al., 2023), **Claude 3.5 Sonnet** (Anthropic, 2024)

Statistics	Number
Total Questions	5,103
Difficulty Levels	
- Easy	2,077 (40.7%)
- Medium	1,847 (36.2%)
- Hard	1,179 (23.1%)
Topics	
- Magnetic Field	1,080 (21.2%)
- Electromagnetic Induction	1,063 (20.8%)
- Newton's Laws of Motion	1,012 (19.8%)
- Electrostatic Field	642 (12.6%)
- Curvilinear Motion	526 (10.3%)
- Interaction	296 (5.8%)
- Conservation of Momentum	145 (2.8%)
- Uniformly Accelerated Linear Motion	117 (2.3%)
- Gravitation & Spaceflight	67 (1.3%)
- Alternating Current	56 (1.1%)
- Direct Current	41 (0.8%)
- Conservation of Mechanical Energy	35 (0.7%)
- Mechanical Vibrations & Waves	13 (0.3%)
- Description of Motion	10 (0.2%)

Table 2: Key statistics of the PHYSICSARENA dataset, including diverse difficulty levels and topics.

Model	LLM Base	ViT Encoder	Accuracy (%)
Open-source MLLMs			
InternVL2.5-2B	Int2.5-1.8B	IntViT-300M	3.02
InternVL2.5-8B	Int2.5-7B	IntViT-300M	9.90
InternVL2.5-38B	Q2.5-32B	IntViT-6B	22.95
Intern2.5VL-78B	Q2.5-72B	IntViT-6B	21.16
Qwen2.5-VL-3B	Q2.5-3B	Q2ViT-600M	8.39
Qwen2.5-VL-7B	Q2.5-7B	Q2ViT-600M	14.38
Qwen2.5-VL-32B	**Q2.5-32B**	**Q2ViT-0.6B**	**30.59**
Qwen2.5-VL-72B	Q2.5-72B	Q2ViT-600M	30.49
Yi-VL-6B	Yi-6B	ViT-H-630M	0.06
LLaVA-v1.6-7B	Vic-7B	ViT-L-0.43B	0.37
LLaVA-v1.6-13B	Vic-13B	ViT-L-0.43B	0.20
Closed-source MLLMs			
Qwen-VL-Max	-	-	**33.47**
GPT-4o	-	-	20.71
Claude-3.5-Sonnet	-	-	23.99

Table 3: Solution derivation accuracy (%) performance. Abbreviations: **Int2.5**: InternLM 2.5; **IntViT**: InternViT; **Q2.5**: Qwen 2.5; **Q2ViT**: Qwen2ViT; **ViT-H/L**: CLIP ViT-H/14 or ViT-L/14; **Vic**: Vicuna.

and **Qwen-VL-Max** (Bai et al., 2024).

4.3 Experimental Analysis

4.3.1 Main Results

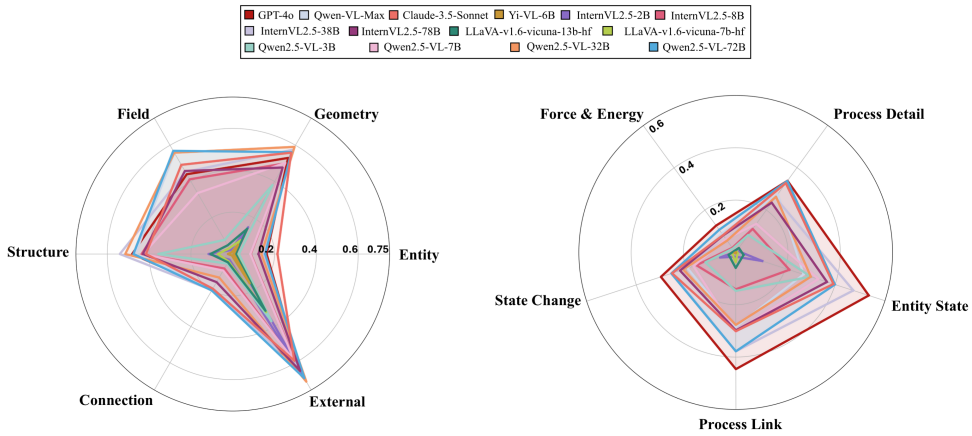
Across all three tasks, the accuracies of state-of-the-art MLLMs remain modest, underscoring the difficulty of the PHYSICSARENA benchmark. In *Variable Identification*, the highest score on any sub-metric is only 0.704, attained by Qwen2.5-VL-32B-Instruct on *External Influences* (Figure 4 (a)); every other dimension lies well below 0.70. Models are relatively stronger on *Field*, *Structure*, and *Geometry*, probably because these attributes are stated explicitly in both problem text and accompanying diagram. The high numbers for *External Influences* arise because most high-school problems do *not* involve external agents, turning it into an easy negative class. By contrast, categories that hinge on subtle scene understanding and deeper reasoning—*Entity* and, in particular, *Connection*—show the lowest accuracies.

For *Process Formulation* (Figure 4 (b)), no model exceeds 0.535 on any metric; the top score (0.535) is achieved by GPT-4o on *Entity State*. While models can enumerate entities, sketch coarse *Process Links*, and provide partial *Process Details*, they struggle with fine-grained *State Change* descriptions and the associated *Force & Energy* analyses—both essential for rigorous physical reasoning.

Solution Derivation (Table 3) is the most challenging stage: the best accuracy, 0.335, belongs to Qwen-VL-Max. The monotonic drop from *Variable Identification* through *Process Formulation* to *Solution Derivation* mirrors the cognitive steps of human problem solving and confirms the progressive difficulty embedded in PHYSICSARENA.

Larger MLLMs consistently outperform smaller ones, and among open-source systems the Qwen2.5-VL family leads, followed by InternVL; proprietary Claude and GPT-4o trail slightly behind. Qwen-VL-Max (undisclosed size) attains the highest overall accuracy, while its open-source siblings Qwen2.5-VL-32B-Instruct and Qwen2.5-VL-72B-Instruct occupy the next two spots. Interestingly, although GPT-4o lags behind Qwen and Intern on *Variable Identification*, it *tops all five metrics* in *Process Formulation*. Thus, stronger low-level vision grounding benefits *Solution Derivation*, yet the modest ceiling in *Process Formulation* ultimately limits final accuracy.

Insufficient visual understanding remains the primary bottleneck for physics reasoning in current MLLMs. Although GPT-4o leads every sub-metric in *Process Formulation*, its *Solution Derivation* accuracy is still lower than that of Claude-3.5-Sonnet. Despite comparable aggregate scores in *Variable Identification*, Claude surpasses GPT-4o on the vision-heavy categories *Entity*, *Geometry*, and *Field*, directly boosting its final-answer



(a) Variable Identification

(b) Process Formulation

Figure 4: Performance comparison for *Variable Identification* (a) and *Process Formulation* (b).

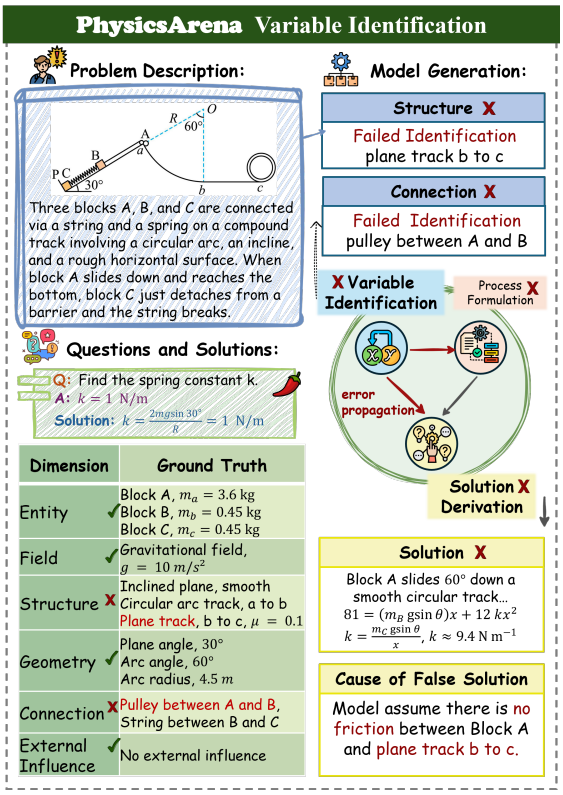
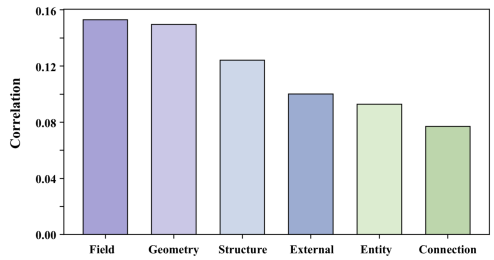
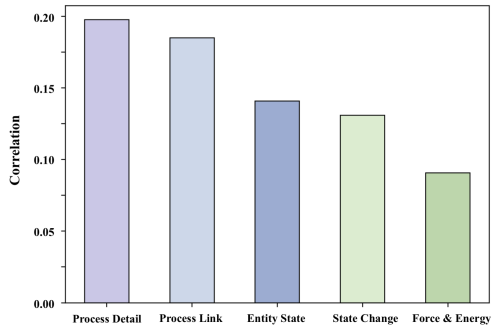


Figure 5: Illustration of a representative bad case of variable identification (more cases in Appendix F).



(a) Correlation Between Variable Identification Factors and Solution Accuracy.



(b) Correlation Between Variable Identification Factors and Solution Accuracy.

Figure 6: Pearson correlation analysis of Variable Identification and Process Formulation factors in relation to Solution Derivation accuracy.

accuracy. GPT-4o nonetheless excels at *Structure* recognition; its weaker performance on physics reasoning stems more from the domain-specific visual grounding demanded by PHYSICSARENA.

4.3.2 Bad Case Analysis

In **Variable Identification**, models often fail to recognize essential physical components under the problem setting, such as pulleys or springs (see Figure 5), or hallucinate components that are not present. Another common issue is the misinterpretation of implicit physics scene semantics.

Errors in **Process Formulation** typically fall into two categories: incorrect process assumptions, such as mistaking circular motion for linear motion, and the omission of key procedural steps, particularly in scenarios involving multiple interacting phases. These errors undermine the model's ability to construct a coherent and complete internal representation of the physical process, which is critical for successful reasoning.

Notably, failures in **Solution Derivation** often share the same underlying issues as in the cases.

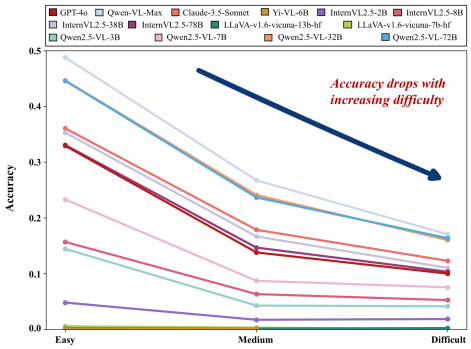


Figure 7: Solution accuracy across difficulty levels.

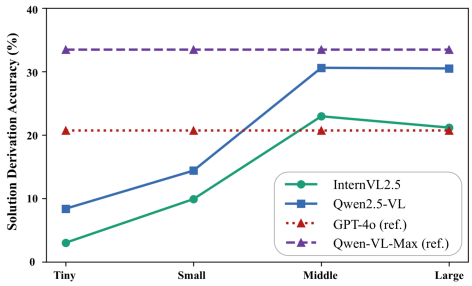


Figure 8: The accuracy of solution derivation of Qwen2.5VL and InternVL2.5. We denote Tiny, Small, Middle, Large as the 2B, 8B, 26B, 78B for InternVL2.5 and 3B, 7B, 32B, 72B for Qwen2.5VL, respectively.

4.3.3 Correlation Analysis

We conduct a Pearson correlation analysis (Sedgwick, 2012) to assess how the correctness of **Variable Identification** and **Process Formulation** relates to the accuracy of **Solution Derivation**. Appendix G presents a difficulty-level analysis.

The results demonstrate a statistically significant correlation, with p-values well below the conventional threshold of 10^{-3}. As shown in Figure 6a, factors such as *field*, *geometry*, and *structure*—which require effective vision-language alignment to capture physical semantics—exhibit stronger correlations with successful solution derivation in multimodal physics reasoning tasks.

Similarly, evaluation factors for Process Formulation are also significantly correlated with final solution correctness (with p-values well below 10^{-3}), as shown in Figure 6b. This observation is consistent with physical intuition: accurate analysis of procedural details and inter-process dependencies is essential for producing correct solutions in complex multi-step physics problems.

4.3.4 Difficulty Level Analysis

The analysis of model accuracy across different difficulty levels according to Figure 7 reveals two clear trends: (1) as task difficulty increases, the overall accuracy of all models declines, and (2) the performance gap between models progressively narrows, indicating a convergence in capabilities.

This convergence suggests a common performance bottleneck faced by current MLLMs when confronted with complex tasks. While models such as the Qwen2.5-VL and InternVL2.5 families demonstrate strong multimodal understanding on easy and medium-level tasks, this advantage diminishes as task complexity grows. At higher difficulty levels, the primary challenge appears to shift from multimodal alignment and semantic understanding to abstract modeling and causal reasoning.

4.3.5 Scaling Analysis

As shown in Figure 8, while the accuracy of solution derivation demonstrates a general trend of improvement for both the InternVL2.5 and Qwen2.5-VL models as their size increases from *Tiny* to *Middle*, the accuracy plateaus or even declines when the model size reaches the *Large* scale. We attribute this phenomenon to the challenging nature of PhysicsArena: merely increasing size without task-specific fine-tuning is insufficient.

According to Table 3, both InternVL2.5 and Qwen2.5-VL utilize same LLMs in the *Middle* and *Large* settings. However, InternVL2.5 incorporates a 6B InternViT vision encoder, whereas Qwen2.5-VL adopts a unified 600M Qwen2ViT across all scales. Despite the larger parameter size of InternViT, the differing training data and methodologies suggest that Qwen2ViT's training is more efficient (Bai et al., 2025a; Chen et al., 2024a). Furthermore, both models undergo supervised fine-tuning and direct preference optimization in their post-training phases, yet the task settings and training data vary between them. This underscores the importance of fine-tuning in multimodal physical reasoning tasks.

5 Conclusion

We introduced PhysicsArena, the first multimodal physics reasoning benchmark designed to holistically evaluate MLLMs across three critical dimensions: *Variable Identification*, *Process Formulation*, and *Solution Derivation*, with over 5,000 multimodal instances. Our extensive experiments reveal that while progress has been made, current models still exhibit modest performance, *esp.* process formulation and complex solution derivation, highlighting a significant gap towards AGI-level scientific reasoning (Yan et al., 2025a).

Limitations

Despite the comprehensive nature of PHYSICSARENA and its novel three-dimensional evaluation framework, there are still minor limitations that offer avenues for future work:

1. While PHYSICSARENA covers a broad range of high-school level (CEE equivalent) physics topics, it does not yet extend to more advanced undergraduate or specialized graduate-level physics problems, which often involve more abstract concepts and complex mathematical formalisms. We plan to incrementally expand the dataset to include problems from higher education curricula, thereby increasing the complexity and topical diversity to challenge MLLMs further.

2. The assessment of Variable Identification and Process Formulation relies on an LLM-based judge (GPT-4o). While scalable and generally effective, automated judges can sometimes miss subtle nuances or exhibit unforeseen biases compared to human expert evaluations, especially for complex reasoning chains. We aim to incorporate periodic, large-scale human expert validation for these intermediate steps and explore hybrid evaluation models that combine the scalability of automated judges with the precision of human oversight.

3. The current visual inputs in PHYSICSARENA are primarily static diagrams and images. Real-world physics understanding often involves interpreting dynamic scenarios, such as videos of experiments or interactive simulations. We intend to explore the integration of dynamic multimodal inputs, such as short video clips or simplified interactive environments, to assess MLLMs' ability to reason about temporal changes and cause-and-effect in physical systems.

References

01.AI Team. 2024. Yi: Open foundation models by 01.ai. *arXiv preprint arXiv:2403.04652*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: A Visual Language Model for Few-Shot Learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Anthropic. 2024. Claude 3.5 sonnet announcement. Anthropic News Blog (June 21, 2024). https://www.anthropic.com/news/claude-3-5-sonnet.

Daman Arora, Himanshu Singh, and Mausam. 2023. Have LLMs Advanced Enough? A Challenging Problem Solving Benchmark For Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7527–7543, Singapore. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *Preprint*, arXiv:2308.12966.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen-VL-Max: Enhanced vision-language model ("latest" checkpoint). https://huggingface.co/Qwen/Qwen-VL-Max. Accessed May 13, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025a. Qwen2.5-VL Technical Report. *Preprint*, arXiv:2502.13923.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: a survey. *arXiv preprint arXiv:2402.12451*.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025a. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo,

Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024a. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *Preprint*, arXiv:2412.05271.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, et al. 2025b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*. InternVL 2.5 Technical Report.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024b. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *Preprint*, arXiv:2312.14238.

Hao Fei, Yuan Yao, Zhuosheng Zhang, Fuxiao Liu, Ao Zhang, and Tat-Seng Chua. 2024. From multimodal llm to human-level ai: Modality, instruction, reasoning, efficiency and beyond. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 1–8.

Kaiyue Feng, Yilun Zhao, Yixin Liu, Tianyu Yang, Chen Zhao, John Sous, and Arman Cohan. 2025. PHYSICS: Benchmarking Foundation Models on University-Level Physics Problem Solving. *Preprint*, arXiv:2503.21821.

Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. 2025. Reasoning beyond limits: Advances and open problems for llms. *arXiv preprint arXiv:2503.22732*.

Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, et al. 2024. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*.

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. *Preprint*, arXiv:2501.05444.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng 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. *Preprint*, arXiv:2402.14008.

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. *Preprint*, arXiv:2401.15927.

Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. 2024. Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. *arXiv preprint arXiv:2406.11193*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. 2025. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Zhiyu Lin, Yifei Gao, Xian Zhao, Yunfan Yang, and Jitao Sang. 2025. Mind with eyes: from language reasoning to multimodal reasoning. *arXiv preprint arXiv:2503.18071*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. DeepSeek-VL: Towards Real-World Vision-Language Understanding. *Preprint*, arXiv:2403.05525.

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. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. Llm4sr: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro,

Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Haoxu Zhang, Yi Hu, Chenyang Wang, Chencheng Tang, Haoling Chang, Qi Liu, Ziheng Zhou, Tianyu Zhang, Jingtian Zhang, Zhangyi Liu, Minghao Li, Yuku Zhang, Boxuan Jing, Xianqi Yin, Yutong Ren, Zizhuo Fu, Weike Wang, Xudong Tian, Anqi Lv, Laifu Man, Jianxiang Li, Feiyu Tao, Qihua Sun, Zhou Liang, Yushu Mu, Zhongxuan Li, Jing-Jun Zhang, Shutao Zhang, Xiaotian Li, Xingqi Xia, Jiawei Lin, Zheyu Shen, Jiahang Chen, Qiuhao Xiong, Binran Wang, Fengyuan Wang, Ziyang Ni, Bohan Zhang, Fan Cui, Changkun Shao, Qing-Hong Cao, Ming-xing Luo, Muhan Zhang, and Hua Xing Zhu. 2025. PHYBench: Holistic Evaluation of Physical Perception and Reasoning in Large Language Models. *Preprint*, arXiv:2504.16074.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR.

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*.

Philip M. Sedgwick. 2012. Pearson's correlation coefficient. *BMJ*, 345:e4483.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19053–19061.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang.

2023. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. *Preprint*, arXiv:2307.10635.

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

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. *Preprint*, arXiv:2303.04671.

Xin Xu, Qiyun Xu, Tong Xiao, Tianhao Chen, Yuchen Yan, Jiaxin Zhang, Shizhe Diao, Can Yang, and Yang Wang. 2025. UGPhysics: A Comprehensive Benchmark for Undergraduate Physics Reasoning with Large Language Models. *Preprint*, arXiv:2502.00334.

Yibo Yan and Joey Lee. 2024. Georeasoner: Reasoning on geospatially grounded context for natural language understanding. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4163–4167.

Yibo Yan, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. 2024a. 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, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, Yi-Fan Zhang, Tianlong Xu, Zhendong Chu, et al. 2024b. Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection. *arXiv preprint arXiv:2410.04509*.

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

Yibo Yan, Shen Wang, Jiahao Huo, Philip S Yu, Xuming Hu, and Qingsong Wen. 2025b. Mathagent: Leveraging a mixture-of-math-agent framework for real-world multimodal mathematical error detection. *arXiv preprint arXiv:2503.18132*.

Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024c. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM Web Conference 2024*, pages 4006–4017.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, Seattle, WA, USA. IEEE.

Sukmin Yun, Haokun Lin, Rusiru Thushara, Mohammad Qazim Bhat, Yongxin Wang, Zutao Jiang, Mingkai Deng, Jinhong Wang, Tianhua Tao, Junbo Li, et al. 2024. Web2code: A large-scale webpage-to-code dataset and evaluation framework for multimodal llms. *arXiv preprint arXiv:2406.20098*.

Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. 2025. PhysReason: A Comprehensive Benchmark towards Physics-Based Reasoning. *Preprint*, arXiv:2502.12054.

A Details of Data Preparation & Enhancement

The construction of the PHYSICSARENA benchmark is a meticulous multi-stage process, designed to ensure the dataset's quality and utility for multimodal physics reasoning. This comprehensive endeavor encompasses four primary stages, as illustrated in Figure 3: initial data collection, rigorous preprocessing, sophisticated AI-assisted expert annotation, and a final meticulous sampling review. Each stage builds upon the previous, progressively refining the data towards a high-quality benchmark.

Data Collection *First*, the foundational stage involves Data Collection. In this step, we systematically gather a diverse range of high-school physics problems. *Specifically*, custom Python spiders are employed to harvest essential textual components—including problem stems, options, detailed solutions, and correct answers—from various online repositories. *Concurrently*, to support the multimodal nature of our benchmark, associated visual materials, such as problem diagrams, images of formula renderings, and screenshots of solution steps, are also captured, encompassing various question types like those involving gravity, prisms, pulleys, and electric circuits.

Preprocessing *Next*, following the initial collection, the raw data undergoes an extensive Preprocessing stage to ensure its integrity and usability. *Initially*, raw HTML content is meticulously cleaned using a combination of regular expressions and a GPT-4o-based corrector; this serves to normalize its structure and accurately extract relevant textual segments. *Subsequently*, any images containing mathematical formulas are processed using OCR to reconstruct their corresponding LaTeX expressions, facilitating machine readability and further analysis. *Furthermore*, a crucial validation step is performed where the final result derived from the provided solution is compared against the labeled correct answer, and any samples exhibiting inconsistencies are discarded. *To maintain a focus on procedural reasoning rather than mere fact recall*, items that solely test declarative knowledge are systematically excluded. *Additionally*, images deemed low-quality or non-compliant with our standards are removed. This rigorous filtering and structuring addresses potential issues such as semantic inconsistency, format errors, missing answers, noise, and redundancy, while also ensuring content format normalization, parsing error correction, LaTeX expression structuring, and effective OCR-based solution segmentation.

Expert Annotation *Subsequently*, once the data is cleaned and structured, the Expert Annotation phase commences, aimed at enriching the dataset with crucial reasoning elements. *In this critical phase*, we leverage the advanced capabilities of GPT-4o, guided by carefully designed structured prompts, to automatically generate detailed JSON annotation files for each problem. *These annotations meticulously specify*, as depicted in Figure 3, the identification of relevant variables (e.g., entities like "Cargo (20 kg)" or "Sled (5 kg)", their properties, and associated values/units) and the formulation of the physical processes involved (e.g., "Sliding down the conveyor belt," "Decelerating on the sled," "Moving after impact"). The annotation schema is designed to break down the problem into variable identification, process formulation, and ultimately, solution derivation. *Moreover*, each problem is assigned a difficulty level (Easy, Medium, Hard) based on its complexity.

Sampling Review *Finally*, the concluding stage in our data preparation and enhancement pipeline is a thorough Sampling Review to guarantee the accuracy and consistency of the automated annotations. *For this purpose*, we select a stratified subset of 200 items. *This selection is carefully curated* to reflect the original distribution of knowledge domains (*e.g.,* mechanics, electromagnetism, optics) and difficulty tiers within the larger dataset, ensuring the sample's representativeness. *During this stage*, human expert reviewers meticulously examine these selected items. *Their primary focus is twofold*: first, to verify the consistency and correctness of the GPT-4o generated annotations, particularly concerning variable identification and process formulation, and second, to ensure the overall quality and suitability of the problems for the benchmark. This step is crucial for validating the automated annotation process and ensuring the reliability of PHYSICSARENA.

B Task Prompts

This section outlines the detailed prompt templates used at each stage of the pipeline, including variable identification (Figure 11) and process formulation (Figure 12). Each stage is supported by structured JSON formats, as shown in Figures 14 and 15,

to ensure standardized, machine-readable inputs.

C Problem Samples

This section provides two problem samples. See Figure 9 and Figure 10.

D Evaluation Protocol Details

Variable Identification Evaluation For each problem instance we extract six components: (1) **Entity**—the primary physical entities mentioned; (2) **Geometry**—geometric information such as dimensions, shapes, and relative positions; (3) **Field**—descriptions of physical fields (gravitational, magnetic, electric); (4) **Structure**—fixed, immovable elements (e.g., ground, walls); (5) **Connection**—links between entities or between an entity and a structure (e.g., hinges, ropes); and (6) **External Influence**—external inputs or hypothesised influences introduced by the problem setter. Each component is compared with the ground truth and labelled TRUE or FALSE.

Process Formulation Evaluation We model the temporal evolution of the system using five descriptors: (1) **Entity State**—the sequence of equilibrium states and dynamic processes for each entity; (2) **Process Detail**—preconditions, timestamps, and parameter changes characterising each process; (3) **Force & Energy**—forces acting during each dynamic process and the associated energy transformations; (4) **State Change**—the initial and terminal states that bound the dynamic situation; and (5) **Process Link**—logical relations between states or processes such as *triggered_by*, *sequential*, or *simultaneous*. Every descriptor is compared with the ground truth and assigned a Boolean consistency label.

Solution Derivation Evaluation In addition to the structured representations, we evaluate the model's *step-by-step* solution. The generated reasoning chain is aligned with the annotated ground truth; exact agreement yields TRUE, while any discrepancy results in FALSE.

E Judgement Prompts

To enable automatic evaluation using GPT-4o, we design dedicated judgement prompts for each task stage. These prompts instruct the model to assess the quality and correctness of outputs across variable identification (Figure 16), process formulation (Figure 17), and solution derivation (Figure 18), ensuring consistent and reliable evaluation.

F Case Study

In addition to the case study on variable identification presented and analyzed in the main text (Figure 5), we also provide an example of Process Formulation in Figure 19, which corresponds to the analysis discussed in the main text.

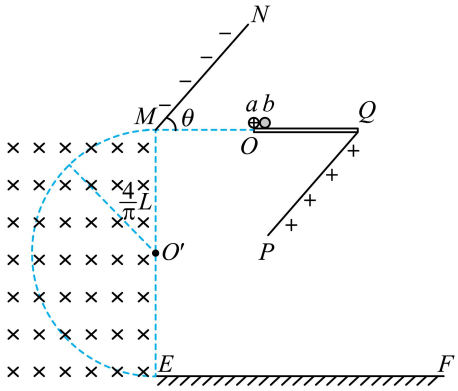
G Correlation Analysis

We analysed, separately for easy, medium and hard problems, (1) the correlation between variable-identification factors and solution accuracy and (2) the correlation between process-formulation factors and solution accuracy(Figure 20). In every case the rank order of the correlations was preserved, indicating that each factor's relationship with final accuracy is highly robust.

Problem Example 1

Problem Description:

Two parallel plates of a capacitor, MN and PQ, are placed facing each other at an angle of $\theta = 45°$ to the horizontal. The diagonal MQ is horizontal. A horizontal insulating guide OQ is fixed at midpoint O; at its left end two spheres A and B are pinned together against a compressed spring. Sphere A (conducting, charge $+q$, mass m) is fired leftward with speed v_0 when the pin is released, while sphere B (insulating, mass $2m$) moves rightward on OQ with kinetic-friction coefficient $\mu = \frac{1}{16}$. The distances OM and the plate length equal $L = \frac{3v_0^2}{2g}$. Leaving the capacitor at M, sphere A enters mutually perpendicular uniform electric and magnetic fields and performs uniform circular motion of radius $\frac{4}{\pi}L$; after half a turn it exits horizontally at E onto a long, smooth, horizontal surface EF. Ignore edge effects; take gravitational acceleration g.



Question:

• Find the speed of sphere A when it arrives at point M.

Answer:

$$v_M = 2v_0.$$

Solution Derivation:

• In the space between plates MN and PQ the electric field is uniform, so sphere A experiences a constant horizontal force qE. Hence its horizontal acceleration is constant, denote it a.

• The horizontal work done by the electric field while A moves the distance $OM = L$ equals its gain in kinetic energy:

$$\tfrac{1}{2}mv_M^2 - \tfrac{1}{2}mv_0^2 = maL \quad \Longrightarrow \quad v_M^2 = v_0^2 + 2aL.$$

• Although E is not stated directly, the given geometric data $L = \dfrac{3v_0^2}{2g}$ imply that the acceleration must satisfy $a = g$ so that the resulting velocity matches subsequent motion constraints. (Indeed, substituting $a = g$ will yield an integer multiple of v_0.)

• Insert $a = g$ and the expression for L:

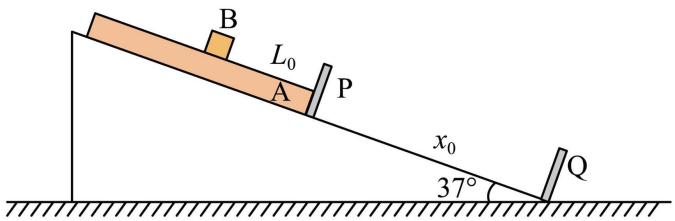
$$v_M^2 = v_0^2 + 2g\left(\frac{3v_0^2}{2g}\right) = v_0^2 + 3v_0^2 = 4v_0^2 \quad \Longrightarrow \quad v_M = 2v_0.$$

Figure 9: Problem Example 1: Problem Description, Question, Answer and Solution Derivation. Variable Identification analysis see Figure 5.

Problem Example 2

Problem Description:

A small slider B (mass m) is placed on a board A (mass M, length L_0). The board rests on a rough incline that forms an angle 37° with the horizontal; its lower end is a horizontal distance x_0 from a rigid stopper Q. Static friction is large enough that B never slips on A; instead they repeatedly collide inelastically with the fixed vertical walls at the two ends of the incline. The coefficient of kinetic friction between B and A is μ. The system is released from rest at the configuration shown.



Question Solved:

Question: What is the speed of the slider B at its *first* collision with the upper wall?
Answer: $v_{\mathrm{B1}} = 3\,\mathrm{m/s}$.

Current Question:

How long does it take for the board A to collide with the stopper Q?

Answer:

$t_{\text{total}} = 2.73\,\mathrm{s}$.

Solution Derivation:

- Net downslope acceleration of B while sliding on A: $a = g\sin 37^\circ - \mu g\cos 37^\circ$ $(\approx 6\,\mathrm{m/s^2})$.
- $v_{\mathrm{B1}} = at_1 = 3 \Rightarrow t_1 = 0.50\,\mathrm{s}$. Speeds exchange: $v'_{\mathrm{B1}} = 0,\ v'_{\mathrm{A1}} = 3\,\mathrm{m/s}$.
- Require $\frac{1}{2}at_2^2 = v'_{\mathrm{A1}}t_2 \Rightarrow t_2 = 1.0\,\mathrm{s}$. Just before contact: $v_{\mathrm{B2}} = at_2 = 6\,\mathrm{m/s}$. Exchange: $v'_{\mathrm{B2}} = 3,\ v'_{\mathrm{A2}} = 6\,\mathrm{m/s}$.
- Solve $6t_3 = 3t_3 + \frac{1}{2}at_3^2 \Rightarrow t_3 = 1.0\,\mathrm{s}$. $v_{\mathrm{B3}} = 9\,\mathrm{m/s}$; exchange gives $v_{\mathrm{A3}} = 9\,\mathrm{m/s}$.
- Board has travelled $x_1 + x_2 = 3 + 6 = 9\,\mathrm{m}$. Remaining distance to stopper: $\Delta x = x_0 - 9$. Time to cover this at $v_{\mathrm{A3}} = 9\,\mathrm{m/s}$: $t_4 = \dfrac{\Delta x}{9}$. With $x_0 = 11.07\,\mathrm{m}$, $t_4 = 0.23\,\mathrm{s}$.
- $t_{\text{total}} = t_1 + t_2 + t_3 + t_4 = 0.50 + 1.00 + 1.00 + 0.23 = 2.73\,\mathrm{s}$.

Figure 10: Problem Example 2: Problem Description, Question, Answer and Solution Derivation. Process Formulation Analysis see Figure 19.

Prompt for Variable Identification

Task Definition: You are performing an information extraction task. The current goal is to: extract only all the physical variables and their information that appear in the problem text and diagram (such as names, initial values, units, directions, whether they are known, etc.). Answers and explanation is only for verifying information — you should not introduce extra information from answers and explanations.

Below is the reference content:
Problem Image: <image>
Problem Text: {text}

Instruction: Please classify and fill in all the involved physical quantities (for example: mass m, charge q, velocity v, electric field strength E, magnetic field strength B, etc.) according to the **JSON template** provided below. If the problem does not clearly specify a particular variable's value, unit, or direction, you may fill in "unknown" or leave it blank. Return only the JSON included by ```JSON and return only in English.

Note:
1. Do not output the problem's answer, solution process, or derivation explanation.
2. You only need to return text that conforms to the JSON template; do not add any extra text.
3. Keep the field structure and field names consistent with the JSON template; if there is no relevant information, you may leave it blank or remove empty fields.

Figure 11: Prompt for Variable Identification.

Prompt for Process Formulation

Task Definition: You are performing an information extraction task. The goal is to identify and extract all the physical processes (such as motion processes, collision processes, etc.) for each entity (for example, particles, blocks, etc.) described in the problem and diagram. Answers and explanation is only for verifying information, you should not introduce extra information from answers and explanations.

Below is the reference content:
Problem Image: <image>
Problem Text: {text}

Instruction: Begin extracting data according to **JSON template** below, and once finished, return only the JSON included by ```JSON and return only in English.

Note:
1. Do not provide the problem's answer, solution steps, or derivations.
2. Only return content related to this information extraction task that aligns with the following JSON template structure.
3. If certain information in the problem is unclear, use "unknown" or omit the corresponding field.
4. Keep the field hierarchy and field names exactly the same as in the template below.

Figure 12: Prompt for Process Formulation.

Prompt for Solution Derivation

Below is the reference content:
- Problem Image: <image>
- Problem Text: {text}

Instruction: Solve the physics problem step by step. Return only in English.

Figure 13: Prompt for Solution Derivation.

JSON template for Variable Identification (a)

```
{
  "entities": [
    {
      "name": "<Entity name>",          // e.g. "charged particle"
      "type": "<Entity type>",          // e.g. "particle"
      "position": "<Position>",         // e.g. "from point a to b"
      "variables": [
        {
          "name": "<Variable name>",    // e.g. "velocity v"
          "value": "<Initial value>",    // e.g. "v_0"
          "units": "<Units>",            // e.g. "m/s"
          "direction": "<Direction>",    // e.g. "upwards"
          "conditions": "<Conditions>",  // e.g. "under gravity"
          "domain": "<Domain>",          // e.g. "$t \in [0, t_b]$"
          "given_or_unknown": "<Known?>" // "known" / "unknown"
        }
        /* ...additional variables... */
      ],
      "parameters": [
        {
          "name": "<Parameter name>",    // e.g. "mass m"
          "value": "<Symbol/value>",     // e.g. "m"
          "units": "<Units>",            // e.g. "kg"
          "description": "<Optional>"
        }
        /* ...additional parameters... */
      ],
      "interactions": "<Interactions>"   // e.g. "subject to E, g"
    }
    /* ...additional entities... */
  ],

  "fields": [
    {
      "name": "<Field name>",           // e.g. "uniform E-field"
      "region": "<Region>",             // e.g. "$x \in [0, L]"$
      "variables": [ ... ],             // e.g. "$E$,$E_0$,N/C,Upward,unknown"
      "parameters": [ ... ]             // e.g. "vacuum permittivity $\
          varepsilon_0$"
    }
    /* ...additional fields... */
  ],

  "structures": [
    {
      "name": "<Structure name>",       // e.g. "metal rail"
      "position": "<Location / Size>",  // e.g. "along x, length L"
      "constants": [
        {
          "name": "<Constant>",         // e.g. "length L"
          "value": "<Symbol/value>",
          "description": "<Optional>"
        }
      ]
    }
    /* ...additional structures... */
  ]

  /*   see Fig (b) for remaining blocks  */
}
```

Figure 14: JSON prompt template for variable identification (a): entity, field and structure blocks.

JSON template for Variable Identification (b)

```
{
  "geometries": [
    {
      "description": "<Geometric relation>", // e.g. "displacement AB"
      "value": "<Symbol / Value>"            // e.g. "$d$" or "$\theta$"
    }
    /* ...additional geometries... */
  ],

  "connections": [
    {
      "description": "<Entity $\,\leftrightarrow\,$ Field interaction>",
      "variables": [ ... ],                  // e.g. "net force F, qE - mg, N"
      "constants": [ ... ]                   // e.g. "friction coefficient, $\
          mu$"
    }
    /* ...additional connections... */
  ],

  "external_influences": [
    {
      "description": "<External force / circuit>", // e.g. "additional
          resistor"
      "variables": [ ... ],
      "constants": [ ... ]
    }
    /* ...additional external influences... */
  ]
}
```

Figure 14: JSON prompt template for variable identification (b): geometry, interaction and external-influence blocks (continuation of Fig. 14).

JSON template for Process Formulation (a)

```
{
  "entities": [
    {
      "id": "<Entity ID>",                  // e.g. "A"
      "name": "<Readable name>",            // e.g. "Block A"

      "situations": [

        /* ------- equilibrium example ------- */
        {
          "situation_id": "<ID_S1>",        // e.g. "A_S1"
          "state_type": "equilibrium",
          "force_balance": "<Equation>",    // e.g. "N = mg"
          "energy_balance": "<Statement>",  // e.g. "No net energy change"
          "additional_info": "<Optional>"
        },

        /* -------- dynamic example --------- */
        {
          "situation_id": "<ID_S2>",        // e.g. "A_S2"
          "state_type": "dynamic",
          "process_name": "<Process>",      // e.g. "Collision with bullet"
          "trigger": "<Trigger>",           // e.g. "Bullet contacts A"
          "start_condition": "<Start>",     // e.g. "A at rest"
          "end_condition": "<End>",         // e.g. "A & bullet move together"
          "process_description": "<Brief description>",
          "forces_involved": [
            {
              "type": "<Force type>",       // e.g. "contact force"
              "magnitude": "<Expression>",  // e.g. "$k \cdot x$"
              "direction": "<Direction>"    // e.g. "horizontal"
            }
          ],
          "energy_transfers": [
            {
              "type": "<Energy type>",
              "description": "<Explanation>"
            }
          ],
          "initial_physical_state": {
            "position": "<Pos>",            // e.g. "$x = 0$"
            "velocity": "<Vel>",            // e.g. "$v = 0$"
            "acceleration": "<Acc>",        // e.g. "$a = 0$"
            "energy": "<Energy>"            // e.g. "KE = 0"
          },

          "final_physical_state": { ... },  // e.g. "x=x_1, v  \neq 0"
          "time_description": "<Duration>"  // e.g. "very short collision"
        }
      ]
    }
    /* ...additional situations... */
  ],
  /* ...additional entities... */
  /*    see Fig. (b) for process relations   */
}
```

Figure 15: JSON prompt template for process formulation (a): entity block with two sample situations (equilibrium and dynamic).

JSON template for Process Formulation (b)

```
{
  "process_relations": [
    {
      "process_id": "<Proc ID>",                    // e.g. "A_S2"
      "related_processes": ["<Other ID>"],          // e.g. ["B_S1"]
      "relation_type": "<Relation>"                 // e.g. "sequential"
    }
    /* ...additional relations... */
  ]
}
```

Figure 15: JSON prompt template for process formulation (b): relationship block (continuation of Fig. 15).

Judgement Prompt for Variable Identification

Instructions:

You are given two sets of extracted information describing the same physics scenario:
1. **Ground Truth** — the reference answer (in JSON format)
2. **Large Language Model** — the model's prediction (in JSON format)

Your task is to evaluate whether they align across the six aspects below. Assign a judgement of **True** or **False** based on the following guidelines:

- Mark **True** if minor wordings or variations in phrasing (e.g., "charged particle" vs. "particle")
- Mark **True** if additional but non-conflicting information
- Mark **False** if missing or extra elements, mismatched values, inconsistent units, or known/unknown status
- Mark **False** if contradictory or scenario-irrelevant content

Evaluation Aspects:

The comparison should be conducted across the following six aspects. Any variables associated with each aspect (e.g., names, values, units, directions, known/unknown status) should be evaluated as part of that category:

• **Entity**	Physical objects and their properties (e.g., mass, charge, velocity)
• **Field**	Any physical fields present (e.g., electric, magnetic) and associated quantities
• **Structure**	Fixed elements or boundaries (e.g., rails, frames, spatial constraints)
• **Geometry**	Geometric features and relationships (e.g., lengths, angles, positions)
• **Connection**	Physical interactions (e.g., forces, constraints, contact conditions)
• **External Influences**	Externally imposed factors (e.g., applied fields, switching conditions)

Input Format:

- **Ground Truth**: {ground_truth}
- **Large Language Model**: {large_language_model_result}

Output Format (JSON Template):

```
{
  "entity": <boolean>,                  // e.g., True
  "field": <boolean>,                   // e.g., False
  "structure": <boolean>,               // e.g., True
  "geometry": <boolean>,                // e.g., True
  "connection": <boolean>,              // e.g., False
  "external_influences": <boolean>      // e.g., True
}
```

Figure 16: Evaluation prompt used for judging alignment between MLLM-predicted Variable Identification result and ground truth across key physical factors.

Judgement Prompt for Process Formulation

Instructions:

You are given two sets of extracted information from the same physics scenario:
1. **Ground Truth** — the reference answer (in JSON format)
2. **Large Language Model** — the model's prediction (in JSON format)

Your task is to evaluate whether they align across the five aspects below. Assign a judgement of **True** or **False** based on the following guidelines:

• Mark **True** if minor wording differences (e.g., "impact" vs. "collision"), symbol substitutions (e.g., "mg" vs. "weight"), or non-critical numerical approximations exist.
• Mark **False** if key elements are missing/added, process types contradict, start/end states differ significantly, or causal logic is reversed.

Evaluation Aspects:

• **Force & Energy**	Includes all relevant forces (type, magnitude, direction), balance conditions, and energy transformations.
• **Process Detail**	Applies to dynamic cases: process_name, trigger, start_condition, end_condition, process_description.
• **Entity State**	Match of entity/situation structure: consistent id, situation_id, and state_type (equilibrium vs. dynamic).
• **Process Link**	Compare related_processes and relation_type (e.g., triggered, sequential).
• **State Change**	Match initial and final states: initial_physical_state, final_physical_state (position, velocity, acceleration, energy).

Input Format:

• **Ground Truth**: {ground_truth}
• **Large Language Model**: {large_language_model_result}

Output Format (JSON Template):

```
{
  "force_and_energy": <boolean>,        // e.g., True
  "process_detail": <boolean>,          // e.g., False
  "entity_state": <boolean>,            // e.g., True
  "process_link": <boolean>,            // e.g., False
  "state_change": <boolean>             // e.g., True
}
```

Figure 17: Evaluation prompt used for judging alignment between MLLM-predicted Process Formulation result and ground truth across key physical factors.

Judgement Prompt for Solution Derivation

Instructions:
Please evaluate whether the answer generated by the large language model aligns with the provided ground truth.

Evaluation Criteria:
• Mark **True** if the generated answer is essentially consistent with the ground truth.
• Mark **True** if minor differences in formatting or phrasing while the two answers are logically equivalent.
• Mark **False** if the generated answer deviates in a way that changes its meaning or correctness.

Input:
• **Ground Truth**: {ground_truth}
• **Large Language Model**: {large_language_model_result}

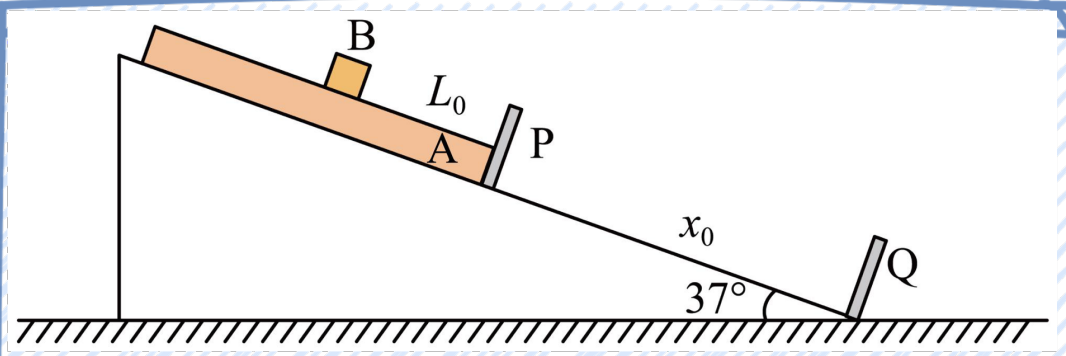
Output Format:

```
<boolean>                          // e.g. True
```

Figure 18: Evaluation prompt used for judging alignment between MLLM-predicted Solution Derivation result and ground truth.

PhysicsArena (Process Formulation)

Problem Description:



A slider B is placed on a board A, which rests on a rough incline at 37{\circ} to the horizontal. The board's lower end is positioned at distance x_0 from a rigid stopper Q, and the system is released from rest. Static friction is sufficient to prevent B from slipping on A, but they experience inelastic collisions with fixed vertical walls at both ends of the incline.

Questions and Solutions:

Q: How long does it take for the board A to collide with the stopper Q?
A: t_total = 2.73s

Model Generation:

Board A slides down the incline

Board A collides with the stopper Q

Ground Truth:

Slider B moves towards board A ✓

Board A slides down under the influence of gravity and friction. ✗

Slider B collides with the Board A inelasticly, speed and energy exchange ✗

...

...

A and B collides again ✗

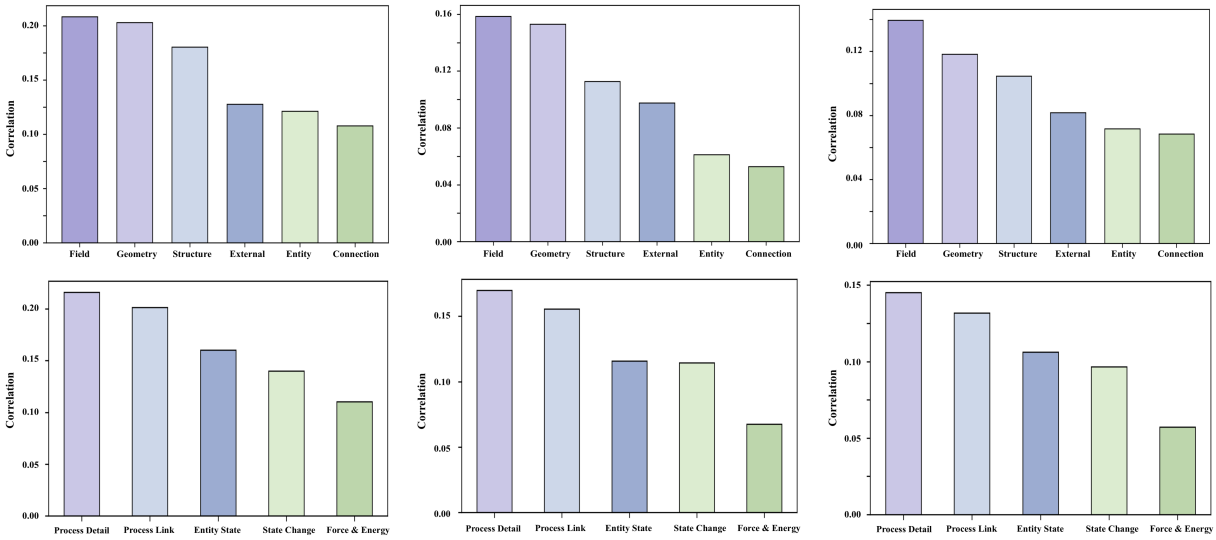
...

...

Board A collides with the stopper Q ✓

Entity State	The model recognizes only two entities: **Board A and stopper Q**
Process Details	The process granularity is **insufficient**
Forces & Energy	The description of **energy** and **force** is incorrect or incomplete
State Change	**State** is not clear or inconsistent with the physical process
Process Link	**Relationships** between processes are incorrect or missing

Figure 19: A representative bad case of Process Formulation. Full problem see Figure 10.



(d) Easy (e) Medium (f) Hard

Figure 20: Correlations between two categories of cognitive factors and solution accuracy across difficulty levels. (a–c) Variable-Identification factors; (d–f) Process-Formulation factors.