PPRBench: A Process-Level Benchmark for Evaluating Multi-Step Physical Reasoning in Large Language Models

Anonymous ACL submission

Abstract

Evaluating whether Large Language Models (LLMs) reason correctly, rather than merely producing correct answers, remains a key challenge in NLP. Existing physics benchmarks measure only outcome accuracy, obscuring whether models perform genuine multi-step reasoning or rely on heuristics. We introduce **PPRBench**, a process-level benchmark for structured physical reasoning, comprising 101 olympiad-level test problems and 3,028 training problems with expert-annotated derivations and scoring rubrics. We also propose **PhysGrader**, an LLM-as-judge framework for rubric-guided evaluation, achieving Pearson r=0.775 agreement with human experts. Across 16 frontier models, we find a substantial reasoning-fidelity gap: most process-level scores remain below 60%, and reasoning models often diverge from explicit canonical derivations. We further identify inadmissible model substitution, where models apply simplified but inappropriate physical frameworks to produce plausible derivations. A small-scale Qwen2.5-1.5B reinforcement-learning pilot suggests that process-level rewards can improve explicit reasoning consistency, but absolute gains remain modest. PPRBench provides the NLP community with a contamination-resistant testbed for diagnosing explicit multi-step structured reasoning. Code and data are available at https://anonymous.4open.science/r/PPRBench_Anonymized.

1 Introduction

A central question in NLP is whether LLMs that produce correct answers are actually reasoning correctly. A model that arrives at the right answer through formula memorization, spurious pattern matching, or by substituting a simpler physical model is not exhibiting the structured multi-step reasoning that benchmarks aim to measure. Yet most evaluation paradigms assess only final-answer correctness, leaving the reasoning process opaque.

Physics problem-solving provides an ideal testbed for this question. Unlike many NLP reasoning tasks, physics problems impose hard constraints: each derivation step must be physically valid, causally grounded, and consistent with the prescribed model. While LLMs have demonstrated strong performance on mathematics (Cobbe et al., 2021; Li et al., 2025; Balunović et al., 2025; Liu et al., 2024b) and coding (Guo et al., 2024; Chen, 2021; Roziere et al., 2023), physics reasoning (Pang et al., 2025; Jaiswal et al., 2024) requires strict fidelity to fundamental laws (Xu et al., 2025b), making it a stringent probe of genuine multi-step reasoning capability.

Current physics benchmarks suffer from two structural limitations (Table 1). First, benchmarks such as PHYBench (Qiu et al., 2025) and ABENCH-PHYSICS (Zhang et al., 2025) assess only final-answer correctness, which cannot distinguish genuine reasoning from shortcut strategies like formula memorization, lucky guessing, or spurious pattern matching all remain invisible to outcome-only evaluation. Second, questions sourced from public textbooks or competitions introduce pretraining data contamination risk (He et al., 2024), making it difficult to assess whether models are reasoning or recalling.

To address these limitations, we introduce **PPRBench**, a process-level benchmark comprising 101 contamination-free expert-curated test problems and 3,028 training problems, each paired with official step-by-step rubrics that decompose solutions into independently assessable reasoning steps. We further propose **PhysGrader**, an LLM-as-judge framework that performs rubric-guided semantic alignment between model-generated reasoning traces and reference derivations, achieving Pearson $r = 0.775$ agreement with human experts. A comparison with existing benchmarks is shown in Figure 1.

Our evaluation of 16 frontier models reveals a

Table 1: Comparison of PPRBench with other benchmarks. For **Level**, HS: High School, GD: Graduate, OP: Olympiad.

Benchmark	Level	Size	Question Type	Scoring Method	Process Evaluation	Data Source
MMLU(Hendrycks et al., 2020)	HS-GD	629	Multiple Choices	Exact Match	✗	Public
SciBench(Wang et al., 2023)	GD	64	Numerical	Numerical Proximity	✗	Public
OlympiadBench(He et al., 2024)	OP	1958	Open-Ended	Rule-Based	✗	Public
ABench-Physics(Zhang et al., 2025)	GD+OP	500	Numerical	Numerical Proximity	✗	Public
PHYBench(Qiu et al., 2025)	HS-OP	500	Expression	EED Score	✗	Private
PHYSICS(Feng et al., 2025)	GD	1297	Expression	Rule + Model	✗	Public
PhysUniBench(Wang et al., 2025)	GD	3308	Fill In/Multiple Choices	Exact Match	✗	Public
SciDA(Zhou et al., 2025)	GD	350	Numerical (Parametric)	Numerical Proximity	✗	Public
PPRBench	Olympiad	101+3028	Expression, Open-Ended	Analytical Equivalence	✓	Private + Public

critical reasoning-fidelity gap: most process-level accuracies fall below 60%, and dedicated reasoning models do not dominate the benchmark, a finding invisible to outcome-only evaluation. Through systematic case analysis, we identify prevalent failure patterns: inadmissible model substitution (45% of errors), restricted solution-class coverage (24%), formula misapplication (19%), and truncated derivations (12%). We also conduct an exploratory Qwen2.5-1.5B training pilot showing that process-level reward signals from PhysGrader can guide reinforcement learning, while also admitting that absolute gains remain small.

Our contributions are:

- **Data Resource**: **PPRBench**, a contamination-resistant benchmark of 101 test and 3,028 training problems, each paired with expert-annotated step-by-step rubrics enabling process-level evaluation, being the first such resource for physics reasoning.
- **Evaluation Framework**: **PhysGrader**, an LLM-as-judge framework for scalable, rubric-guided semantic alignment of multi-step reasoning traces, achieving high human agreement and providing process-level reward signals for training.
- **Model Analysis & Interpretability**: Systematic evaluation of 16 frontier LLMs uncovering the reasoning-fidelity gap, reasoning drift in reasoning models, and a taxonomy of four failure modes that outcome-only metrics cannot detect.
- **NLP Engineering Experiment**: A small-scale GRPO pilot on Qwen2.5-1.5B showing modest improvements in step-by-step reasoning consistency, reported as feasibility evidence rather than a claim of broad training generalization.

To avoid data contamination, the test set remains hidden from the public to maintain the integrity of the long-term leaderboard. Researchers can submit their models for performance verification.

2 PPRBench: Benchmarking Process-Level Physics Reasoning

In this section, we (i) formulate physics solutions as structured reasoning trajectories, (ii) construct and annotate problems via a controlled pipeline (§ 2.2), (iii) propose PhysGrader for process-level semantic alignment (§ 2.3), and (iv) validate the benchmark with statistical analyses (§ 2.4).

2.1 Task Formulation

For each problem p, we formalize the reference solution as a structured reasoning trajectory $\mathcal{T}_p = \{\tau_1, \tau_2, \ldots, \tau_n\}$, where each τ_i denotes a critical intermediate derivation, symbolic transformation, or conceptual milestone. This trajectory serves as the basis for fine-grained process-level evaluation.

2.2 Data Collection and Quality Assurance

The construction of PPRBench proceeds via a four-stage pipeline, as illustrated in Figure 3, which systematically transforms each physics problem p and its reference solution into a rigorously validated benchmark instance.

Stage 1: Multi-source Problem Acquisition We begin by collecting candidate problems p from elite physics competitions and advanced pedagogical materials. Each problem undergoes rigorous expert verification to ensure it necessitates complex multi-step reasoning, introduces novel physical scenarios, and supports fine-grained evaluation. To strictly mitigate data contamination, test instances are exclusively sourced from non-public Chinese examination papers and are manually audited against public web indices. We preserve the original Chinese problem statements and prompts throughout evaluation; this reduces translation artifacts and contamination risk, but it also means that cross-model comparisons partly reflect Chinese-language physics reasoning ability rather than language-independent physics competence.

Table 2: Scores of LLMs on PPRBench. Domain abbreviations are MC: Mechanics, EM: Electromagnetism, TD: Thermodynamics, OT: Optics, TR: Theory of Relativity, and QP: Quantum/Modern Physics. O denotes Final-Answer Accuracy (outcome-level) and P denotes the PhysGrader score (process-level). Bold and underline mark the highest and second-highest scores.

Models	**MC**		**EM**		**TD**		**OT**		**TR**		**QP**		**Overall**	
	O	P	O	P	O	P	O	P	O	P	O	P	O	P
Proprietary LLMs														
Gemini-2.5 Flash (Comanici et al., 2025)	51.61	53.83	41.17	43.86	**68.75**	47.85	66.67	68.79	40.00	52.82	66.67	71.91	55.45	54.82
Gemini-2.5 Pro (Comanici et al., 2025)	**64.52**	63.89	47.06	51.10	43.75	51.67	**75.00**	**70.76**	40.00	50.46	66.67	68.16	57.43	58.88
Gemini-3 Pro (DeepMind)	61.29	**69.41**	64.71	**63.52**	37.50	**59.47**	58.33	64.75	30.00	57.86	66.67	**72.74**	55.45	**65.11**
Claude Opus 4.1 (Anthropic)	29.03	40.17	29.41	28.32	56.25	27.20	33.33	52.23	20.00	44.72	66.67	51.88	38.61	39.09
Claude Sonnet 4.5 (Anthropic, 2025)	32.26	40.57	29.41	42.46	31.25	32.85	25.00	62.23	20.00	40.72	40.00	54.39	30.69	45.85
GPT-4o (Hurst et al., 2024)	12.90	14.01	29.41	8.07	18.75	9.14	16.67	30.71	10.00	5.25	46.67	15.40	21.78	13.44
GPT-5 (Singh et al., 2025)	54.84	60.82	**70.59**	47.87	50.00	47.82	66.67	65.70	50.00	**64.79**	**86.67**	69.27	**62.38**	57.76
OpenAI o3 (OpenAI)	41.94	51.81	47.06	43.31	50.00	41.48	33.33	57.88	**60.00**	53.59	46.67	59.38	45.54	50.25
OpenAI o4 (OpenAI)	38.71	52.45	41.18	35.74	37.50	45.58	58.33	55.63	50.00	45.27	73.33	55.30	47.52	46.71
Grok 4 (xAI)	48.39	52.33	52.94	33.44	37.50	28.89	33.33	57.16	30.00	40.99	66.67	47.02	46.53	42.50
Open-source LLMs														
kimi-k2 (Team et al., 2025)	38.71	45.90	41.18	29.46	31.25	37.04	25.00	57.09	30.00	39.86	40.00	56.68	35.64	42.57
Deepseek-v3 (Liu et al., 2024a)	25.81	42.34	29.41	28.44	37.50	44.24	50.00	56.14	30.00	59.98	60.00	52.05	36.63	43.80
Qwen2.5-1.5B (Qwen et al., 2025)	6.45	1.05	11.76	5.36	0.00	2.53	0.00	0.83	10.00	1.50	0.00	1.67	4.95	2.12
Qwen3-8B (Team, 2025)	9.68	7.25	5.88	7.34	6.25	0.63	0.00	0.00	0.00	0.00	0.00	7.19	4.95	4.69
Qwen3-32B (Team, 2025)	12.90	17.09	35.29	11.86	25.00	11.53	25.00	17.58	20.00	21.25	26.67	12.03	22.77	14.51
Qwen3-235B-A22B-instruct (Team, 2025)	48.39	58.19	52.94	44.22	56.25	36.13	50.00	69.26	50.00	62.73	**86.67**	69.22	56.44	55.12
Tuning Strategies														
Qwen2.5-1.5B, SFT	/	1.69	/	3.99	/	2.17	/	0.42	/	0.96	/	3.48	/	2.18
Qwen2.5-1.5B, GRPO	/	2.28	/	3.39	/	4.05	/	0.28	/	3.80	/	3.17	/	2.79

Stage 2: Precision Digitization (OCR) For each selected problem p and its accompanying reference solution S_p, scanned documents are converted into a machine-readable representation $(\tilde{p}, \tilde{S}_p)$ using the MinerU (Niu et al., 2025) OCR pipeline. Annotators then conduct symbol-level verification of all mathematical expressions to correct OCR-induced errors and ensure fidelity of notation.

Stage 3: Reasoning Trajectory Formalization and Rubric Synthesis The digitized solution $\tilde{S}_p$ is further decomposed into a structured reasoning trajectory $\mathcal{T}_p = \{\tau_1, \tau_2, \ldots, \tau_n\}$, where each τ_i denotes an atomic reasoning step, encompassing intermediate derivations, symbolic transformations, or conceptual milestones. Each step inherits its score from the official grading rubric, yielding a process-annotated instance $(p, \mathcal{T}_p)$ equipped with normalized scoring weights. When official answers provide multiple valid solution paths, we preserve these alternatives as explicitly marked rubric branches with corresponding scores. To guarantee logical coherence and mitigate reasoning gaps, these trajectories undergo rigorous expert verification.

Stage 4: Consensus-based Final Review Finally, each annotated instance $(p, \mathcal{T}_p)$ undergoes independent double review by expert annotators. Any discrepancies are resolved through senior adjudication, resulting in a finalized benchmark entry approved through consensus.

Table 3: Disciplinary statistics in PPRBench. MC: Mechanics, EM: Electromagnetism, TD: Thermal dynamics, OT: Optics, QP: Quantum Physics, TR: Theory of Relativity.

Domain	MC	EM	TD	OT	QP	TR	Total	Source
#Test	31	17	16	12	15	10	101	Exam papers
#Train	846	431	504	398	571	278	3028	Books

Step-level weights are derived directly from official scoring rubrics prepared by physics-competition instructors, not assigned heuristically. For each problem p, we convert the raw rubric scores a_i for step τ_i into normalized weights $w_i = a_i / \sum_j a_j$, preserving relative importance while making the scores comparable between the problems. Annotators only transform the teacher-provided rubric into our structured format; they do not invent new weights.

2.3 LLM Evaluation Framework: PhysGrader

Process-level evaluation requires verifying the semantic alignment between model-generated reasoning traces and reference derivations. Unlike outcome-based verification, this assessment is intractable via exact string matching or rigid symbolic rules, as valid reasoning steps often exhibit significant syntactic diversity despite being semantically equivalent. Manual grading, while accurate, is prohibitively expensive and difficult to scale across thousands of problems and models. To enable scalable, consistent, and semanti-

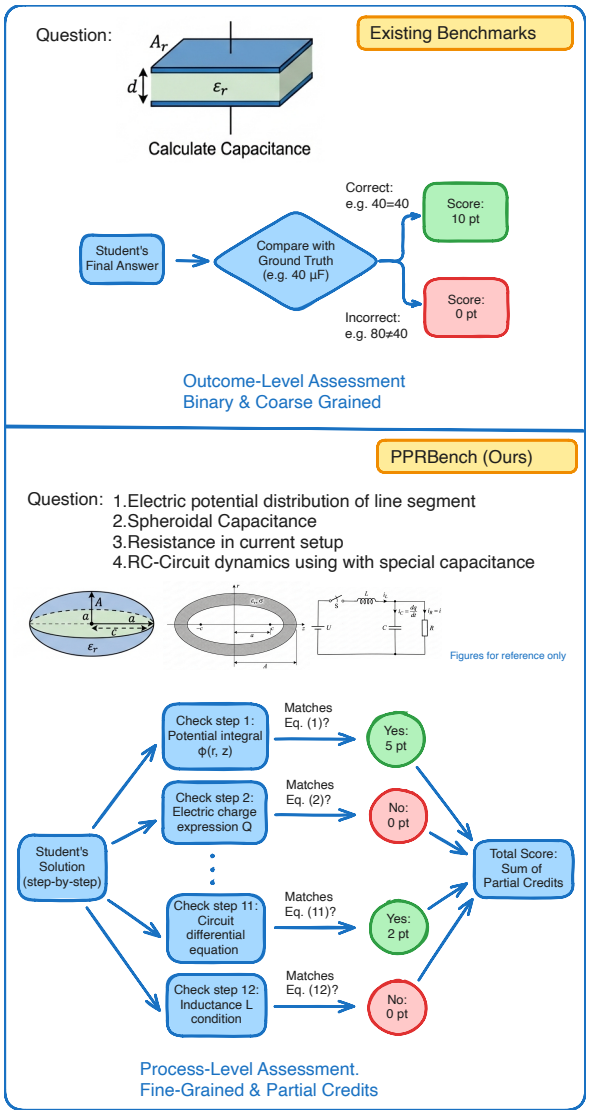

Figure 1: Comparison between PPRBench and previous work.

cally aware evaluation, we develop **PhysGrader**, an LLM-as-a-judge framework (Gu et al., 2024; Zheng et al., 2023; Gilardi et al., 2023; Cheng et al., 2025) designed to assess alignment between model-generated reasoning trajectories and reference ones.

Given a problem p with annotated reasoning trajectory $\mathcal{T}_p = \{\tau_1, \tau_2, \ldots, \tau_n\}$, and a model-generated trajectory $\hat{\mathcal{T}}_p$, PhysGrader evaluates whether each reference step τ_i is explicitly reflected in $\hat{\mathcal{T}}_p$. Formally, we define a binary alignment indicator $\delta(\hat{\mathcal{T}}_p, \tau_i) \in \{0, 1\}$, which equals 1 if the model-generated trajectory contains an expression, derivation, or logical statement semantically equivalent to τ_i, and 0 otherwise. Each step τ_i is associated with a normalized weight w_i, reflecting its conceptual importance. The final score for problem p is computed as

$$S(p) = \sum_{i=1}^{n} w_i \, \delta(\hat{\mathcal{T}}_p, \tau_i),$$

and overall model performance is obtained by averaging $S(p)$ across all evaluated problems.

This definition deliberately measures *explicit rubric-aligned process fidelity*: credit is assigned only when a required physical step is present or semantically recoverable from the submitted trajectory. As a consequence, a terse answer that reaches the correct final expression but omits intermediate derivations can be underscored. We therefore interpret PhysGrader scores as evidence about displayed reasoning traces, not as a complete measure of a model's latent physics ability. The efficacy of this structured process-level evaluation protocol is illustrated in Figure 5.

2.4 Data Analysis

PPRBench comprises a training set of 3,028 problems and a verified test set of 101 problems across six physics domains (Table 3), with Mechanics representing the largest share (31%), followed by Electromagnetism and Thermodynamics. Figure 4 shows the length distribution of questions and answers, evidencing the its complexity. The test set is sourced exclusively from non-public materials and manually verified to prevent contamination.

Evaluation Reliability We conducted a correlation analysis against human expert ratings across $n = 101$ samples. Using our *Reasoning Chain Formalization* as ground truth, the Pearson correlation between PhysGrader (the average of Qwen3-235B-A22B-instruct and GPT-5) and the average of seven human experts reached 0.775. This marks a substantial improvement over the 0.264 achieved using unstructured reference answers (Figure 5).We use Qwen3-235B-A22B-instruct as the default judge because it offered the best trade-off between symbolic-equivalence accuracy and computational cost in our pilot validation.

Because LLM-as-judge systems can exhibit positional and verbosity biases, PhysGrader is implemented as a reference-step alignment task rather than a pairwise preference task: the judge receives one candidate answer, one rubric item, and the relevant reference step, then decides semantic equivalence. This design removes pair-order effects, but it does not eliminate verbosity sensitivity. A full cross-judge ablation over multiple grader models

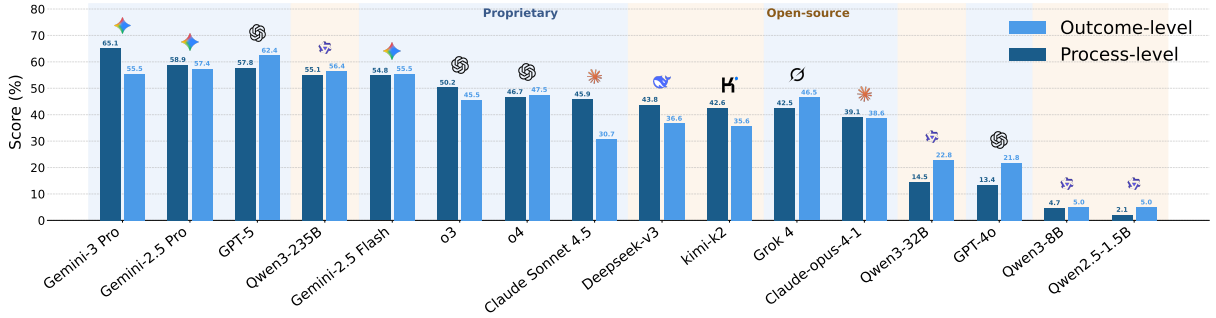

Figure 2: Model performance on PPRBench. The ranking is based on process-level scores.

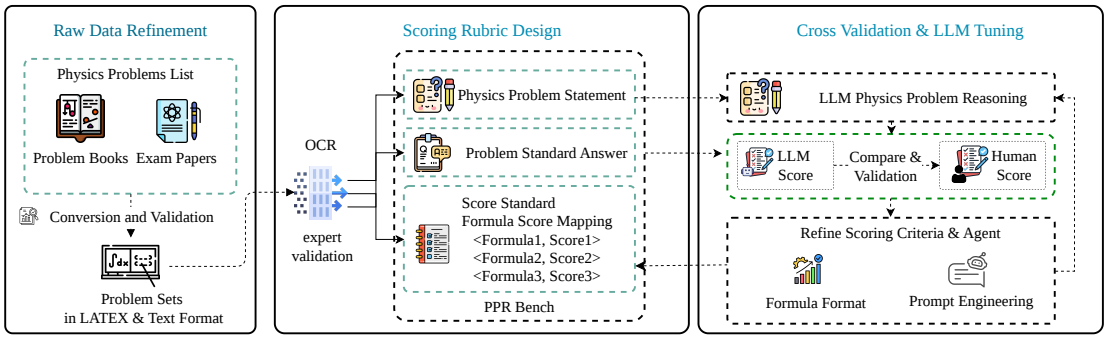

Figure 3: **Data construction pipeline of PPRBench.** The pipeline transforms raw physics materials into a rigorously validated benchmark through Raw Data Refinement, Scoring Rubric Design (incorporating process-level annotations), and Cross Validation, utilizing a feedback loop between LLM outputs and human expert verification to ensure high-fidelity evaluation.

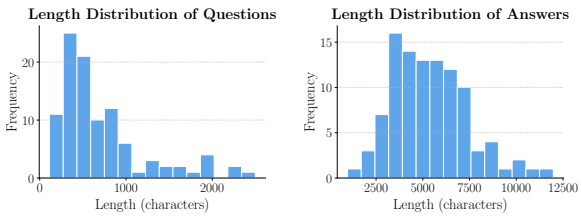

Figure 4: Distribution of question length and answer length in PPRBench.

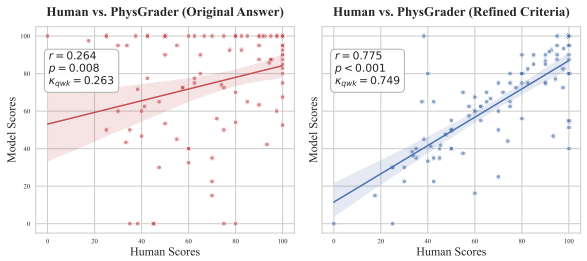

Figure 5: Correlation analysis between human ratings and model-generated scores. Left: LLM grading based on original answer; right: based on refined rubric.

and rubric orderings remains future work; consequently, all reported process scores should be read as Qwen3-235B-A22B-instruct PhysGrader scores rather than judge-invariant ground truth.

3 Experiments

We evaluated 16 LLMs: 10 proprietary models and 6 open-weight models. For proprietary APIs, we sample one response per prompt using default parameters. Open-weight models are hosted locally on eight 80GB NVIDIA A100 GPUs with temperature $= 0.2$ and top_p $= 0.95$. All problems and prompts are in Chinese, so results should be interpreted as Chinese-language physics reasoning performance.

3.1 Main Findings

We report Final-Answer Accuracy ("Outcome-level") and PhysGrader Score ("Process-level"), averaged over all samples (Table 2, Figure 2). All models use the *same* user prompt template (Appendix A), without model-specific system prompts or few-shot demonstrations, though provider defaults and Chinese-language preferences may still affect comparisons.

PPRBench challenges state-of-the-art models Process-level scores vary widely, mostly between 10% and 60%. Gemini-3 Pro reaches 65.11%, whereas GPT-4o reaches only 13.44%, showing

that general language fluency does not ensure robust physics reasoning. The spread also indicates that PPRBench separates models by displayed derivational fidelity rather than only by generic instruction-following ability.

Reasoning-oriented models falter on PPRBench Dedicated reasoning models do not dominate PPRBench: o3 and o4 rank 6th (50.25%) and 7th (46.71%), behind high-capacity general models such as Gemini-3 and GPT-5. This suggests that self-correction and longer search do not automatically satisfy strict physical constraints, and may even produce "reasoning drift" over long traces. In other words, more deliberation helps only when each intermediate step remains tied to the correct physical model.

Final-Answer accuracy doesn't translate into reasoning quality High Final-Answer Accuracy (FAA) does not guarantee valid reasoning. Several models with competitive FAA hallucinate physical laws or make unsupported intermediate leaps, suggesting shortcut-based correctness rather than robust derivation. This decoupling motivates process-level evaluation: a plausible final expression can hide invalid assumptions, missing conservation laws, or algebraic jumps.

Model Scale and Reasoning Effort Scaling generally helps but with diminishing returns: the Qwen 32B-to-235B jump remains insufficient for the hardest problems. Extra reasoning effort is useful only when the trace stays physically consistent, so future progress likely requires both stronger base competence and better trace management.

3.2 Statistical Scope and Uncertainty

PPRBench is intended primarily as a high-difficulty diagnostic benchmark rather than a large-scale leaderboard. The test set contains 101 problems, and domain-level analyses are substantially smaller: for example, Relativity contains 10 problems and Optics contains 12. Since PhysGrader scores are bounded in $[0, 1]$, a conservative normal-approximation 95% confidence half-width for a domain mean is at most $1.96/(2\sqrt{n})$. This yields an overall upper-bound half-width of 9.8 percentage points, but domain-level upper bounds range from 17.6 to 31.0 points. We therefore use domain results to make hypotheses about model behavior, instead of fine-grained statistically conclusive claims.

3.3 Mechanism-Oriented Performance Analysis

Two mechanisms explain the most salient performance gaps among large-scale models: some fluent models substitute the required physical framework with a simpler one; and long reasoning traces can drift away from the canonical derivation even in dedicated reasoning models.

3.3.1 Physical-Model Substitution: Fluent Solutions Can Solve the Wrong Problem

GPT-4o(Hurst et al., 2024) illustrates a different failure mechanism: low process-level performance can arise even when outputs are fluent, structured, and formula-rich. Its 13.44% process-level score is therefore not primarily explained by response length or degeneration. Instead, manual inspection shows that GPT-4o frequently replaces the prescribed physics with a simpler framework, or restricts a problem to a narrower solution class than the prompt requires. We provide detailed case studies and aggregate error-pattern analysis in Appendix C.

The conclusion is that surface-level completeness is not sufficient evidence of physical reasoning fidelity. GPT-4o often supplies many steps for the wrong physical model rather than merely skipping derivations. Outcome-only evaluation can miss this behavior when the final expression is dimensionally plausible, whereas process-level rubrics expose the mismatch between the generated framework and the reference framework.

3.3.2 Reasoning-Trace Drift: Longer Deliberation Can Accumulate Errors

Longer reasoning is beneficial only when the trace remains physically aligned. O3(OpenAI), despite being a dedicated reasoning model, scores **14.9 percentage points** below Gemini-3 Pro(DeepMind) on the *rate* metric (0.502 vs. 0.651) and 9.9 points below on *FAArate* (0.455 vs. 0.554). We term this mechanism **reasoning drift**: an extended reasoning trace diverges from the canonical solution path, accumulating errors that a more concise, structured solver avoids.

Category-level analysis. Table 4 breaks down the gap by physics domain. The observed deficit is largest in Mechanics (+20.2 pp), Optics (+18.0 pp), and Electromagnetism (+17.6 pp)—domains that require multi-step derivations with auxiliary conserved quantities. Relativity has the smallest

Category	n	O3	G3	Δ	CI
Mechanics (M)	31	0.433	0.635	**+0.202**	±0.176
Optics (O)	12	0.415	0.595	**+0.180**	±0.283
Electromagnetism (E)	17	0.518	0.694	**+0.176**	±0.238
Thermodynamics (T)	16	0.594	0.727	+0.133	±0.245
Quantum/Modern (Q)	15	0.579	0.648	+0.069	±0.253
Relativity (R)	10	0.536	0.579	+0.043	±0.310
Overall	**101**	**0.502**	**0.651**	**+0.149**	±0.098

Table 4: Per-category *rate* comparison between O3 and Gemini-3 Pro (G3). Δ = G3 − O3; positive values indicate G3 advantage. CI gives the conservative 95% half-width upper bound for a bounded mean in that category, so small-domain gaps are descriptive rather than conclusive.

observed gap, consistent with its fixed, well-defined solution templates (Lorentz transforms, four-vectors), but this observation is based on only 10 problems and should be treated as exploratory.

Mechanism manifestations. Qualitative analysis of the 38 cases where G3 leads by >0.2 shows how reasoning drift appears in practice, with additional miscellaneous errors detailed in Appendix D:

1. **Mid-chain path deviation.** O3 starts with a correct framework but introduces a non-standard auxiliary quantity (e.g., effective potential in the wrong form), breaking the derivation chain (e.g., ID 6, variable-mass rocket; O3 = 0.15, G3 = 1.00).

2. **Unused conservation law.** O3 identifies a conserved quantity but fails to deploy it strategically in subsequent steps (e.g., ID 19, charged particle in crossed fields; O3 = 0.35, G3 = 1.00).

3. **Step elision.** O3 answers are on average 2,218 characters versus G3's 5,167. Skipped intermediate steps are not matched by the rubric-guided PhysGrader evaluation, causing silent score loss even when the final answer is correct.

These manifestations support a more precise conclusion than "O3 is weak at physics": the gap reflects a *style mismatch* and trace-management problem. O3's high-level, skip-step reasoning is misaligned with a scoring rubric that rewards explicit intermediate formulae, and its longer traces can introduce auxiliary definitions that later derail the solution. This interpretation is intentionally limited: it identifies a mismatch between O3's displayed solution style and our explicit rubric, but it does not prove that O3 lacks the latent reasoning ability to fill in omitted derivations.

3.3.3 Treatment of Alternative Solution Paths

Physics problems often admit more than one valid derivation strategy, such as different coordinate systems, equivalent conservation-law formulations, or alternative algebraic routes to the same physical conclusion. PPRBench therefore does not require every answer to follow a single canonical derivation. For problems with multiple solution methods, the accepted alternative derivations decided by expert teachers are explicitly marked in the reference answer and are accompanied by their corresponding scoring criteria.

During rubric construction, each officially provided alternative is retained as an acceptable branch for the relevant sub-question or reasoning step. In grading, PhysGrader grants credit when a model-generated step is semantically equivalent to any listed branch under the same rubric item, rather than only to the first displayed derivation. Formally, if a scoring item i has a set of acceptable alternatives $\mathcal{A}_i = \{\tau_i^{(1)}, \ldots, \tau_i^{(m)}\}$, then the alignment indicator is positive when

$$\exists k \in \{1, \ldots, m\}, \quad \hat{\mathcal{T}}_p \equiv \tau_i^{(k)}.$$

This design preserves the official competition-style grading standard: alternative approaches receive credit when they are specified by the answer key and satisfy the same physical requirements, while unsupported shortcuts or solutions on a different physical model are not treated as valid alternatives.

3.4 Training with PPRBench

We explored SFT and GRPO as an underpowered feasibility probe on Qwen2.5-1.5B, not as evidence of generalization to larger model families. SFT uses learning rate 1×10^{-5} for 10 epochs, AdamW with cosine scheduling, warmup ratio 0.1, and effective batch size 64 across eight NVIDIA-A100 GPUs. GRPO uses actor learning rate 1×10^{-6}, maximum prompt/response length 8192, KL regularization, and PhysGrader rewards for 300 steps on 8 GPUs (28 hours).

Enhancement strategies are limited in effectiveness. SFT did not yield observable gains. GRPO improved the success rate from 2.12% to 2.79%, a 31.6% relative but only 0.67 percentage-point absolute gain; the critic score rose from 0.0228 to 0.1275. Thus, the reward signal can drive optimization, but the gain is too small to support claims of practical utility.

Patterns observed in post-training model outputs. We observed two behavioral shifts:

- SFT and GRPO responses became longer and more notation-heavy, but this "professionalism" did not consistently improve correctness.
- The baseline and SFT models often skipped intermediate algebra or integrals, while GRPO more often produced detailed subproblem solutions, suggesting that process rewards can encourage step-by-step reasoning.

Overall, the pilot shows that PhysGrader rewards are technically usable, but the results should not be read as a practical training recipe. Larger models, stronger baselines, and ablations separating verbosity from physical correctness are needed before drawing conclusions about training value.

4 Related Work

4.1 Physics Reasoning Benchmarks

Early physics reasoning benchmarks such as MMLU (Hendrycks et al., 2020), SciBench (Wang et al., 2023), and JEEBench (Arora et al., 2023) primarily assess high-school to undergraduate-level problems, typically using multiple-choice or numerical questions. As state-of-the-art reasoning models already excel on these benchmarks, subsequent efforts have explored more challenging settings by increasing problem difficulty or diversity. OlympiadBench(He et al., 2024), PHYBench (Qiu et al., 2025), and UGPhysics (Xu et al., 2025a) incorporate open-ended or expression-based questions that require multi-step derivations. PhoPile (Zheng et al., 2025b) collects problems from Physics Olympiads including the International Physics Olympiad (IPhO). Other benchmarks such as PhysUniBench (Wang et al., 2025) and SeePhys (Xiang et al., 2025) expand scale or introduce multimodal inputs. Basic information of these benchmarks are shown in Table 1. However, all of these benchmarks assess only Final-Answer Accuracy (FAA), a coarse-grained metric that cannot distinguish genuine reasoning from shortcut strategies. PPRBench addresses this gap by providing step-level rubrics that make the reasoning process directly evaluable.

4.2 Process-Level Evaluation

PRM800K (Lightman et al., 2023) introduced Process-level Reward Models (PRMs) for process supervision; Math-Shepherd (Wang et al., 2024) automated this pipeline. ProcessBench (Zheng et al., 2025a) and PRMBench (Song et al., 2025) benchmark PRMs on error localization and error-type diversity. PPRBench extends this line of work to physics, where constraints on valid steps are stricter and more domain-specific than in mathematics.

4.3 LLM-as-Judge Evaluation

The LLM-as-judge paradigm (Gu et al., 2024; Zheng et al., 2023; Gilardi et al., 2023; Cheng et al., 2025) enables scalable evaluation of open-ended outputs but is susceptible to positional bias and holistic inconsistency (Gilardi et al., 2023). PhysGrader addresses this by grounding the judge in an explicit formula-level rubric derived from official scoring criteria, reducing evaluation to a structured alignment task analogous to reference-based NLP metrics (ROUGE, BERTScore) but operating over symbolic reasoning steps. This design reduces, but does not remove, known LLM-judge risks such as verbosity sensitivity.

5 Conclusion

We introduced PPRBench, a process-level benchmark for physical reasoning, and PhysGrader, an LLM-as-judge framework that achieves $r = 0.775$ agreement with human experts. Evaluating 16 frontier models reveals a substantial explicit reasoning-fidelity gap: most models remain below 60% process-level accuracy, and chain-of-thought models can exhibit reasoning drift invisible to outcome-only evaluation. We identify *inadmissible model substitution*, where models substitute a simpler framework for the prescribed one, accounting for 45% of GPT-4o's errors—alongside other failure modes including formula misapplication and truncated derivations. A small Qwen2.5-1.5B training pilot shows that process-level rewards are usable for optimization, but the 0.67 percentage-point absolute GRPO gain should be interpreted as preliminary feasibility evidence rather than a claim of practical training utility. PPRBench contributes a contamination-resistant benchmark, a human-calibrated process-level evaluation framework, and empirical evidence that outcome-level metrics can mischaracterize displayed model reasoning, with possible extensions to other domains requiring explicit multi-step derivations.

6 Limitations

PPRBench's test set contains 101 problems distributed across six physics domains, which limits the statistical power of domain-level comparisons; results for smaller domains such as Theory of Relativity (10 problems) and Optics (12 problems) should be interpreted as exploratory. PhysGrader measures explicit rubric-aligned reasoning traces, so models whose reasoning is correct but terse (e.g., O3) may be systematically underscored; future work should empirically separate latent reasoning ability from rubric-style and verbosity alignment. We use Qwen3-235B-A22B-instruct as the default judge, but we do not yet provide a full cross-judge ablation over multiple grader models, rubric orderings, or verbosity-controlled answers, leaving residual LLM-as-judge bias as a limitation. Our fine-tuning experiments are conducted solely on Qwen2.5-1.5B due to computational constraints, and the absolute GRPO gain is only 0.67 percentage points, so these results should be read as a feasibility pilot rather than evidence of practical training utility or generalization to larger models. All problems and prompts are in Chinese, which reflects the source materials and reduces translation-induced noise, but it may limit direct comparability with English-language benchmarks and may partially confound physics capability with language-of-reasoning preferences. Finally, although all models receive the same user prompt template, provider-specific API defaults and hidden system behaviors may still introduce minor cross-model comparability confounds.

References

Anthropic. System card addendum: Claude opus 4.1. https://www-cdn.anthropic.com/9fa30625273bafdf5af82c93719d7ca606485a16.pdf.

Anthropic. 2025. System card: Claude sonnet 4.5.

Daman Arora, Himanshu Singh, and 1 others. 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.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. 2025. Matharena: Evaluating llms on uncontaminated math competitions. *arXiv preprint arXiv:2505.23281*.

Mark Chen. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. Elephant: Measuring and understanding social sycophancy in llms. *arXiv preprint arXiv:2505.13995*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Google DeepMind. Gemini 3 pro model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

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. *arXiv preprint arXiv:2503.21821*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *The Innovation*.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, and 1 others. 2024. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Raj Jaiswal, Dhruv Jain, Harsh Parimal Popat, Avinash Anand, Abhishek Dharmadhikari, Atharva Marathe, and Rajiv Ratn Shah. 2024. Improving physics reasoning in large language models using mixture of refinement agents. *arXiv preprint arXiv:2412.00821*.

Xiaoyuan Li, Moxin Li, Wenjie Wang, Rui Men, Yichang Zhang, Fuli Feng, and Dayiheng Liu. 2025. Mathopeval: A fine-grained evaluation benchmark for visual operations of mllms in mathematical reasoning. *arXiv preprint arXiv:2507.18140*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024b. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. *arXiv preprint arXiv:2405.12209*.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, Guang Liang, Rui Zhang, Wenzheng Zhang, Yuan Qu, Zhifei Ren, Yuefeng Sun, Yuanhong Zheng, Dongsheng Ma, and 42 others. 2025. Mineru2.5: A decoupled vision-language model for efficient high-resolution document parsing. *Preprint*, arXiv:2509.22186.

OpenAI. Openai o3 and o4-mini system card. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf.

Xinyu Pang, Ruixin Hong, Zhanke Zhou, Fangrui Lv, Xinwei Yang, Zhilong Liang, Bo Han, and Changshui Zhang. 2025. Physics reasoner: Knowledge-augmented reasoning for solving physics problems with large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11274–11289.

Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Haoxu Zhang, Yi Hu, and 1 others. 2025. Phybench: Holistic evaluation of physical perception and reasoning in large language models. *arXiv preprint arXiv:2504.16074*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, and 1 others. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

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

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, and 150 others. 2025. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Lintao Wang, Encheng Su, Jiaqi Liu, Pengze Li, Peng Xia, Jiabei Xiao, Wenlong Zhang, Xinnan Dai, Xi Chen, Yuan Meng, and 1 others. 2025. Physunibench: An undergraduate-level physics reasoning benchmark for multimodal models. *arXiv preprint arXiv:2506.17667*.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *Preprint*, arXiv:2312.08935.

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. *arXiv preprint arXiv:2307.10635*.

xAI. Grok 4 model card. https://data.x.ai/2025-08-20-grok-4-model-card.pdf.

Kun Xiang, Heng Li, Terry Jingchen Zhang, Yinya Huang, Zirong Liu, Peixin Qu, Jixi He, Jiaqi Chen, Yu-Jie Yuan, Jianhua Han, and 1 others. 2025. Seephys: Does seeing help thinking?–benchmarking vision-based physics reasoning. *arXiv preprint arXiv:2505.19099*.

Xin Xu, Qiyun Xu, Tong Xiao, Tianhao Chen, Yuchen Yan, Jiaxin Zhang, Shizhe Diao, Can Yang, and Yang Wang. 2025a. Ugphysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. *arXiv preprint arXiv:2502.00334*.

Yinggan Xu, Yue Liu, Zhiqiang Gao, Changnan Peng, and Di Luo. 2025b. Physense: Principle-based physics reasoning benchmarking for large language models. *Preprint*, arXiv:2505.24823.

Yiming Zhang, Yingfan Ma, Yanmei Gu, Zhengkai Yang, Yihong Zhuang, Feng Wang, Zenan Huang, Yuanyuan Wang, Chao Huang, Bowen Song, and 1 others. 2025. Abench-physics: Benchmarking physical reasoning in llms via high-difficulty and dynamic physics problems. *arXiv preprint arXiv:2507.04766*.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025a. Processbench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1009–1024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Shunfeng Zheng, Yudi Zhang, Meng Fang, Zihan Zhang, Zhitan Wu, Mykola Pechenizkiy, and Ling Chen. 2025b. Benchmarking foundation models with retrieval-augmented generation in olympic-level physics problem solving. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21927–21956.

Junting Zhou, Tingjia Miao, Yiyan Liao, Qichao Wang, Zhoufutu Wen, Yanqin Wang, Yunjie Huang, Ge Yan, Leqi Wang, Yucheng Xia, and 1 others. 2025. Scida: Scientific dynamic assessor of llms. *arXiv preprint arXiv:2506.12909*.

A Prompt Templates

All experiments are conducted in Chinese; the templates below are English translations provided for readability. The logical structure and instructions are identical in both languages. Three templates are documented: the **inference template** (§A.1), the **PhysGrader evaluation template** (§A.2), and the **rubric-synthesis template** (§A.3).

A.1 Inference Prompt Template

The placeholder `{PROBLEM}` is replaced with the full problem statement including all sub-questions. No model-specific system prompt or few-shot demonstration is added.

Inference Prompt Template

You are a physics expert. Please read the following questions and provide a clear, step-by-step solution.

[User]
Answer the following questions:
{PROBLEM}

A.2 PhysGrader Evaluation Template

PhysGrader receives two inputs: the structured rubric produced by the rubric-synthesis template (§A.3) and the model's response. Rubric formulas are identified by sequential tags `tag{1}`, `tag{2}`, ... inherited from the synthesis step. PhysGrader outputs a per-question breakdown and a final score enclosed in `\boxed{}`.

PhysGrader Evaluation Template

[System]
You are a physics teacher. Your task is to grade the [Student Answer] based on the [Scoring Rubric]. Do not add or deduct points on your own.

Input Format
You will receive two parts:

1. **[Scoring Rubric]:** Contains formula tags(e.g., `tag{1}`, `tag{2}`, `tag{3}`), formulas (or original expressions), and point values (e.g., 1 point, 2 points).
2. **[Student Answer]:** The student's response, including text and formulas.

Grading Rules

1. Grade each formula by comparing it against the rubric one by one.
   - If the student's answer contains a formula identical to, physically equivalent to, or dimensionally equivalent to one in the rubric, award the corresponding points.
   - Do not award points more than once for a formula already credited in the student's answer.
   - Do not award points again for a rubric formula already credited.
   - Equivalence criteria: algebraic transformations, symbol adjustments, and substitutions are allowed, but the logic must be fully equivalent and physical dimensions must match. Dimensionally non-equivalent formulas receive no points.
2. Partial credit mechanism.
   - If a sub-question requires multiple formulas for full marks, the student must write all of them.
   - If only some formulas are correct, award only the corresponding partial points.
   - No bonus points for derivations not listed in the rubric.
3. Strictly follow the scoring rubric.
   - Do not award extra points based on solution length or expression complexity.
   - If the student's formula is completely wrong or missing, no points are awarded.
4. Text descriptions do not substitute for formulas.
   - If the student only uses verbal description without writing a formula, no points are awarded.
   - If the verbal description clearly expresses a physical relationship equivalent to a formula (e.g., "the equipotential surface is a prolate spheroid"), the corresponding points may be awarded.

Output Format
Please output the following structure:

1. **[Per-question Scoring]:** List the scoring for each sub-question in order, stating which formulas were correctly identified and the corresponding points awarded. Use "Eq. ①", "Eq. ②" to correspond to the tags in the rubric. If a formula does not appear, mark it as "not present."

2. **[Total Score]:** Sum up the scores for all sub-questions and state the student's final score. Use `\boxed{}` to display the result.

Output Example (do not copy the point values)

[Per-question Scoring]
Sub-question (1) (5 points):
– The student wrote an integral expression for $\varphi(r, z)$, equivalent to Eq. ① $\rightarrow$ 5 points awarded.
Sub-question (2) (14 points):
– The student stated the equipotential surface is a prolate spheroid $\rightarrow$ 4 points awarded.
– Eq. ② not present $\rightarrow$ 0 points.
– Expression equivalent to Eq. ③ found $\rightarrow$ 4 points awarded.
– Expression equivalent to Eq. ④ found $\rightarrow$ 4 points awarded.
Subtotal: 12 points.
[Total Score]
Final score: $5 + 12 + \cdots = \boxed{30}$.

[User]
Grade from scratch using the rubric below.
[Scoring Rubric]:
{RUBRIC}

[Student Answer]:
{MODEL_RESPONSE}

A.3 Rubric-Synthesis Template

During dataset construction, this template decomposes a digitised reference solution into the structured ⟨formula, score⟩ rubric consumed by PhysGrader. Each formula is assigned a sequential tag `tag{`n`}` that serves as its identifier throughout the grading pipeline. All outputs are reviewed and corrected by human experts before being committed to the benchmark.

Rubric-Synthesis Template

[System]
You are a physics teacher. Given a reference solution with point allocations, extract a structured scoring rubric.

Input Format
You will receive:

1. **[Reference Solution]:** Contains sub-question identifiers (e.g., (1), (2), (3), (1.1), (1.2)), formulas with sequential tags `tag{1}`, `tag{2}`, ..., and point values.

Extraction Rules

1. Identify each sub-question and all its tagged formulas.

2. Preserve the original formula exactly as written in the reference solution, including the trailing `tag{`n`}`. Do not split a formula that contains spaces into multiple entries.

3. Record the point value associated with each formula tag.

Output Format
[Per-question Rubric]: List sub-questions in order. For each sub-question, list: formula index, original formula (with trailing `tag{`n`}`), and point value.

Output Example (do not copy the point values)

[Per-question Rubric]
Sub-question (1) (5 points):
Formula 1, $y = a + b$ `tag{1}`, 2 points
Formula 2, $y = a + b$ `tag{2}`, 3 points
Sub-question (2) (14 points):
Formula 1, $y = a + b$ `tag{3}`, 10 points
Formula 2, $y = a + b$ `tag{4}`, 4 points

[User]
Start fresh using the reference solution below; ignore all previous data.
[Reference Solution]:
{REFERENCE_SOLUTION}

B Appendix: Sample Entry from the Physics Benchmark Dataset

Problem (ID: 1)

Question (40 points):

Uniformly Charged Line Segment and Prolate Spheroidal Capacitor

1. Consider a uniformly charged line segment of length $2c$ with linear charge density λ. Taking the midpoint of the segment as the origin and the segment direction as the z-axis in a cylindrical coordinate system, find the electric potential distribution in space.

2. Consider a prolate spheroidal capacitor (a prolate spheroid is obtained by rotating an ellipse about its major axis). The inner electrode is a prolate spheroid with semi-major axis a and focal distance c; the outer electrode is a prolate spheroid with semi-major axis A and focal distance c, where $A > a > c$. The major axes and centers of the two electrodes coincide, and the space between them is filled with a dielectric of relative permittivity ϵ_r. Find the capacitance of this capacitor.

3. Given that the filling dielectric has electrical conductivity σ, find the resistance between the inner and outer electrodes.

4. The capacitor above is equivalent to a pure capacitor in parallel with a pure resistor. This equivalent RC unit is connected in series with an inductor L, a DC source (constant EMF, negligible internal resistance), and a switch to form a closed circuit. Initially the switch is open; at some moment it is suddenly closed. Under what condition on L does the system reach steady state as fast as possible?

Useful formula:

$$\int \frac{\mathrm{d}x}{\sqrt{x^2+a^2}} = \ln\left(x+\sqrt{x^2+a^2}\right) + C.$$

Scoring Rubric (Total: 40 points)

Part (1) (5 points):

- Correct derivation of the potential integral and closed-form result (5 pts):

$$\phi(r,z) = \frac{\lambda}{4\pi\epsilon_0} \ln \frac{z+c+\sqrt{r^2+(z+c)^2}}{z-c+\sqrt{r^2+(z-c)^2}} \quad \text{(Eq. 1)}$$

Part (2) (14 points):

- Recognizing that the equipotential surfaces of the charged segment are prolate spheroids, and that the segment's field can substitute for the capacitor's field (4 pts).

- Total charge on the inner electrode:

$$Q = 2\lambda c \qquad \text{(Eq. 2; 2 pts)}$$

- Potentials of inner and outer electrodes (Eq. 3; 4 pts):

$$\phi_{\text{in}} = \frac{\lambda}{4\pi\epsilon_0\epsilon_r} \ln \frac{a+c}{a-c},$$
$$\phi_{\text{out}} = \frac{\lambda}{4\pi\epsilon_0\epsilon_r} \ln \frac{A+c}{A-c}.$$

- Capacitance:

$$C = \frac{Q}{\Delta\phi} = \frac{8\pi\epsilon_0\epsilon_r c}{\ln \dfrac{(a+c)(A-c)}{(a-c)(A+c)}} \quad \text{(Eq. 4)}$$

(4 pts)

Part (3) (8 points):

- Microscopic Ohm's law: $\vec{j} = \sigma\vec{E}$ (Eq. 5; 2 pts)

- Current through a closed surface enclosing the inner electrode:

$$I = \iint_S \vec{j}\cdot \mathrm{d}\vec{S} = \sigma \iint_S \vec{E}\cdot \mathrm{d}\vec{S} = \frac{\sigma Q}{\epsilon_0\epsilon_r} \quad \text{(Eq. 6)}$$

(2 pts)

- Resistance:

$$R = \frac{\Delta\phi}{I} = \frac{\epsilon_0\epsilon_r}{\sigma C} = \frac{\ln \dfrac{(a+c)(A-c)}{(a-c)(A+c)}}{8\pi\sigma c} \quad \text{(Eq. 7)}$$

(4 pts)

Part (4) (13 points):

- KCL: $i_L = \dfrac{\mathrm{d}q}{\mathrm{d}t} + i$ (Eq. 8; 2 pts)

- KVL: $L\dfrac{\mathrm{d}i_L}{\mathrm{d}t} + \dfrac{q}{C} = U$ (Eq. 9; 2 pts)

- Ohm's law for RC branch: $\dfrac{q}{C} = iR$ (Eq. 10; 2 pts)

- Combined ODE:

$$\frac{\mathrm{d}^2 q}{\mathrm{d}t^2} + \frac{1}{RC}\frac{\mathrm{d}q}{\mathrm{d}t} + \frac{q}{LC} = \frac{U}{L} \quad \text{(Eq. 11)}$$

(3 pts)

- Critical damping condition (fastest approach to steady state):

$$L = 4R^2C = \frac{\epsilon_0\epsilon_r}{2\pi\sigma^2 c}\ln\frac{(a+c)(A-c)}{(a-c)(A+c)} \quad \text{(Eq. 12)}$$

(4 pts)

Reference Solution

Part (1). Setting up cylindrical coordinates (r, ϕ, z) with origin at the midpoint of the segment, the potential at point $P(r, z)$ is obtained by integrating over the segment:

$$\phi(r,z) = \frac{\lambda}{4\pi\epsilon_0}\int_{-c}^{c}\frac{\mathrm{d}z'}{\sqrt{r^2+(z-z')^2}} = \frac{\lambda}{4\pi\epsilon_0}\ln\frac{z+c+\sqrt{r^2+(z+c)^2}}{z-c+\sqrt{r^2+(z-c)^2}}$$

Part (2). The equipotential surfaces of the charged segment satisfy

$$\frac{z^2}{K^2} + \frac{r^2}{K^2-c^2} = 1,$$

which are prolate spheroids sharing focal distance c. By the uniqueness theorem, the field of the segment is equivalent to that of the capacitor (with permittivity ϵ_r). The charge is $Q = 2\lambda c$, and the capacitance follows from the potential difference between the two spheroidal surfaces, yielding Eq. (4).

Part (3). Using $\vec{j} = \sigma\vec{E}$ and Gauss's law over a surface enclosing the inner electrode gives $I = \sigma Q/(\epsilon_0\epsilon_r)$, so $R = \epsilon_0\epsilon_r/(\sigma C)$, yielding Eq. (7).

Part (4). The equivalent circuit is an LC loop with a parallel RC load. Applying Kirchhoff's laws yields the second-order ODE (11). The characteristic equation is

$$s^2 + \frac{1}{RC}s + \frac{1}{LC} = 0.$$

Critical damping (fastest non-oscillatory approach to steady state) requires the discriminant to vanish:

$$\frac{1}{R^2C^2} - \frac{4}{LC} = 0 \implies L = 4R^2C.$$

Substituting the expressions for R and C gives Eq. (12).

C GPT-4o Performance Investigation

This appendix expands the brief discussion in Section 3.3.1. GPT-4o achieves a notably low process-level score of 13.44%, substantially below other frontier models such as Gemini-3 Pro (65.11%) and GPT-5 (57.76%). To understand whether this reflects an evaluation artifact or a genuine capability gap, we conducted a comparative case analysis between GPT-4o and GPT-5 on the same 101 test problems.

Comparable Surface-Level Output, Divergent Physical Substance. Surface metrics indicate that GPT-4o produces well-structured, formula-rich outputs: average response length is 1,867 characters (vs. 2,202 for GPT-5; only 18% shorter), and the number of intermediate equations is comparable across difficulty tiers. The performance gap therefore cannot be attributed to truncated or visibly degenerate generations. Instead, the divergence emerges at the level of *which physics is invoked*: PhysGrader records an average of 12.6 missing reference equations per problem for GPT-4o, versus 7.8 for GPT-5. Final-answer accuracy (FAA) shows a similar gap (21.78% vs. 62.38%), and the FAA→process gap is only 3.7 percentage points larger than the FAA gap itself, indicating that the process-level evaluation is consistent with, rather than artificially amplifying, outcome-level differences.

Case Study 1: Inadmissible Physical Approximation. The problem asks for the landing speed, angle, and time of flight of an object launched tangentially at altitude h above Earth, using the exact gravitational potential $-GMm/r$ and an integral identity provided in the prompt for evaluating the time-of-flight integral. GPT-5 correctly applies energy and angular-momentum conservation in the inverse-square field, sets up the radial equation of motion $(\mathrm{d}r/\mathrm{d}t)^2 = v_0^2 - 2\mu/r_0 + 2\mu/r - r_0^2v_0^2/r^2$, and uses the supplied integral formula to obtain a closed-form flight time, scoring 95%. GPT-4o instead replaces the central-force problem with a *flat-Earth, constant-g projectile* model, deriving $t = \sqrt{2h/g}$, $v_f = \sqrt{v_0^2 + 2gh}$, and $\tan\theta = \sqrt{2gh}/v_0$. While these formulas are dimensionally correct and superficially plausible, they answer a different physical problem; PhysGrader's grading report explicitly identifies the substitution as "not equivalent" to the reference, and the problem scores 0%. Notably, this failure mode is not visible

from response length (GPT-4o produced a longer response than GPT-5 on this item) and would be invisible to outcome-only evaluation that accepts approximately correct numerical magnitudes.

Case Study 2: Restricted Solution Class. On the first sub-question ("conditions for periodic motion"), GPT-4o equates centripetal force with gravity to derive a single circular-orbit speed $v = \sqrt{GM/(R+h)}$. The reference solution requires the full elliptic-orbit condition $\sqrt{2GMR/[(R+h)(2R+h)]} \leq v_0 < \sqrt{2GM/(R+h)}$, derived from the bound-orbit energy condition $E < 0$ together with the non-impact constraint enforced via angular momentum conservation. GPT-4o's answer characterizes only a single circular trajectory and misses both the upper (escape) and lower (impact) bounds, receiving 0 of the 15 points for this sub-question. The error is not algebraic but conceptual: the model failed to recognize that "periodic motion" admits the entire family of bound elliptic orbits.

Case Study 3: Truncated Relativistic Derivation. The problem asks for the equation of an expanding sphere ($x^2 + y^2 + z^2 = u^2t^2$) under a Lorentz boost, and the physical meaning when $u = c$. GPT-5 obtains the complete prolate-spheroid form with explicit center offset x_c, semi-axes a and b, and the limiting wavefront $x'^2 + y'^2 + z'^2 = c^2t'^2$, scoring 100%. GPT-4o begins correctly with $R(t) = ut$ and the time transformation $t = t'/\gamma + vx/c^2$ but then truncates mid-substitution, never completing the algebraic reduction to the spheroid form and never discussing the $u = c$ limit. This is the error mode closest to "step skipping," but our analysis shows it is the exception rather than the rule (only 7 of 101 problems exhibit this pattern of complete-failure-with-correct-start).

Aggregate Error Patterns. Manual analysis of all 101 grading reports identifies four dominant failure modes:

- **Inadmissible model substitution (45%):** GPT-4o replaces the prescribed physical model with a simpler one (e.g., constant-g for inverse-square gravity, circular orbit for elliptic, plane wave for prolate spheroid). Grading reports flag these as "not equivalent" (63 of 101 problems contain this verdict at least once) or use "approximation" that the reference framework does not permit (29 problems).
- **Restricted solution class (24%):** The model finds one valid case (e.g., one orbit, one regime, one boundary condition) but misses the full set required by the problem statement.
- **Truncated derivation (12%):** The derivation begins correctly but terminates before reaching the required result, often with an incomplete algebraic step left dangling.
- **Formula misapplication (19%):** Correct formulas are invoked outside their domain of validity (e.g., applying the lensmaker equation in a non-paraxial regime).

Implications. GPT-4o's low process-level score is not an evaluation artifact: response length, formula density, and FAA all corroborate the process-level finding. Rather, GPT-4o exhibits a distinct *model-substitution* failure mode in which the system replaces the problem's prescribed physical framework with a more elementary one it can solve confidently, producing a complete-looking derivation that answers a different problem. This pattern is the inverse of "step skipping"—GPT-4o supplies many steps, but for the wrong physical model. Outcome-only evaluation, which would reward dimensionally-plausible answers, systematically misses this failure mode; PPRBench's process-level rubrics expose it directly.

D Reasoning Drift: Case Studies

This appendix provides concrete evidence for the four reasoning-drift failure modes identified in Section 3.3.2. All cases compare O3 and Gemini-3 Pro (G3) on the same problem; scores are stepwise normalized.

D.1 Failure Mode 1: Mid-Chain Path Deviation

Problem ID 6 — Variable-Mass Rocket in Gravity (Category M, 40 pts). A rocket ejects gas radially at speed u with $dm/dt = -\alpha m$. The three sub-questions ask: (1) whether circular orbit exists and at what radius; (2) orbital stability under small perturbation; (3) second cosmic velocity at radius R.

	O3	G3
rate	0.15	1.00
Machine score	6 / 40	40 / 40

Correct solution path (used by G3):

1. Meshchersky equation $\Rightarrow$ effective radial thrust αu (outward).
2. Circular orbit condition:

$$v = \sqrt{GM/r - \alpha u r}$$

3. Effective potential $V_{\text{eff}}(r)$ with the $\alpha u r$ term $\Rightarrow$ stability via $V''_{\text{eff}} > 0$.
4. Second cosmic velocity from energy conservation with the modified potential.

O3 deviation: O3 correctly derives the circular orbit speed (sub-question 1, partial credit) but then introduces an effective potential of the form

$$V_{\text{eff}}(r) = -\frac{GM}{r} - \frac{\alpha u r^2}{2} + \frac{L^2}{2r^2},$$

where the sign of the αu term is inverted relative to the correct expression. This causes the stability analysis (sub-question 2) to yield the wrong conclusion, and the error propagates to sub-question 3.

Analysis. The initial framework is correct; the drift occurs at the auxiliary-quantity construction step. In a long reasoning chain, a single sign error in an intermediate definition silently invalidates all downstream conclusions—a compounding effect absent in G3's shorter, more modular derivation.

D.2 Failure Mode 2: Unused Conservation Law

Problem ID 19 — Charged Particle in Radial Electric + Uniform Magnetic Field (Category E, 40 pts). A particle of mass m, charge $+q$ orbits at radius R with angular velocity ω in field $E = kr$ and perpendicular B. Sub-question (1): after an impulsive radial kick of ωR, find the maximum radius. Sub-question (2): stability of the circular orbit and radial oscillation period.

	O3	G3
rate	0.35	1.00
Machine score	14 / 40	40 / 40

Key conserved quantity:

$$\mathcal{L} \equiv mr^2\dot{\theta} - \tfrac{1}{2}qBr^2 = \mathcal{L}_0 \qquad (\star)$$

G3: Derives ($\star$) in the first paragraph, uses it to eliminate $\dot{\theta}$ from the energy integral, constructs the effective potential, and reads off r_{max} and the oscillation frequency directly.

O3: Writes $J = mr^2\dot{\theta} + \frac{1}{2}qBr^2$ (equivalent form, different sign convention) but then switches to a Lagrangian approach without substituting $J = \text{const}$, leaving the energy integral in two unknowns. Sub-question (1) is solved only partially; sub-question (2) is not completed.

Analysis. O3 *identifies* the conserved quantity but does not *deploy* it as a reduction tool. This reflects a gap between pattern recognition and strategic planning within the reasoning chain—the model recognises the structure but does not exploit it.

D.3 Failure Mode 3: Step Elision

Problem ID 52 — Fresnel Lens Design (Category O, 40 pts). The problem asks for the sphere radius R_k and sphere-centre-to-focus distance for each annular zone of a Fresnel lens, given refractive index n, focal length f, and zone spacing d.

	O3	G3
rate	0.125	0.90
Machine score	5 / 40	36 / 40

Scoring rubric (sub-question 1, 27 pts): Five key formulae must appear explicitly: (i) Snell's law $\sin\beta = n\sin\alpha$; (ii) sine rule in $\triangle$CAF; (iii) Pythagorean relations for AF and R; (iv) the $h \ll R$ small-angle approximation; (v) the closed-form expression for R_k.

O3: Sets up the correct geometric model and jumps directly to the final expression for R_k, skipping the intermediate sine-rule and Pythagorean steps. The scorer cannot match the absent intermediate formulae, awarding only 5 of 27 points for sub-question 1.

G3: Writes out each geometric relation in sequence, explicitly labelling angles and distances, so every rubric formula is present.

Quantitative evidence. Across all 101 problems, O3 produces answers averaging **2,218 characters** versus G3's **5,167 characters** (ratio 0.43). The Pearson correlation between answer length and *rate* is $r = 0.31$ for O3 ($p < 0.01$), indicating that brevity is associated with lower scores under this rubric.

D.4 Counter-Example: O3 Outperforms G3

Problem ID 69 — Rigid Rod Pendulum Internal Forces (Category M, 40 pts). A uniform rod of length l, mass m is hinged at one end and released from horizontal. Find the cross-sectional force at distance x from the hinge when the rod makes

angle θ with the vertical, and compute the work done by internal forces on a segment $[x, x + dx]$ during an infinitesimal rotation $d\theta$.

	O3	**G3**
rate	0.825	0.225
Machine score	33 / 40	9 / 40

O3: Correctly interprets θ as a general angle, applies energy conservation to obtain $\omega^2 = 3g\cos\theta/l$, differentiates to get $\alpha = -(3g/2l)\sin\theta$, and isolates the cross-sectional force by treating the distal segment as a sub-body.
G3 error: Misreads "the rod makes angle $d\theta$ with the vertical" as a small-angle approximation ($\theta \approx 0$), computing $\omega^2 = 2\alpha\, d\theta$ from kinematics alone. This yields a dimensionally inconsistent expression that matches none of the rubric formulae.
Analysis. This counter-example illustrates that G3's verbose, structured style can also fail when the initial physical interpretation is wrong. O3's more careful reading of the problem statement is an advantage here. The asymmetry—O3 wins on 9 problems, G3 on 38—nonetheless indicates that O3's failure modes are more prevalent under this benchmark's explicit-step scoring protocol.

D.5 Summary

Mode	**Description**	**Est. share**
Path deviation	Mid-chain sign/form error	$\sim$35%
Unused conserved qty	Identifies but doesn't use	$\sim$25%
Step elision	Skips intermediate steps	$\sim$25%
Other	Arithmetic / misc.	$\sim$15%

Table 5: Estimated contribution of each reasoning-drift mode to O3's score deficit relative to G3. Shares are approximate and based on manual inspection of the 38 problems where Δrate > 0.2.

The common thread is a *misalignment between reasoning style and scoring format*: O3 often presents fewer intermediate formulae, whereas PPRBench's rubric-guided PhysGrader protocol rewards explicit derivation steps. Together, these findings suggest that evaluating reasoning models on process-level benchmarks requires disentangling *reasoning ability* from *reasoning-style alignment with the rubric*.