PHYSICSEVAL: Inference-Time Techniques to Improve the Reasoning Proficiency of Large Language Models on Physics Problems

Anonymous ACL submission

Abstract

The discipline of physics stands as a cornerstone of human intellect, driving the evolution of technology and deepening our understanding of the fundamental principles of the cosmos. Contemporary literature includes some works centered on the task of solving physics problems—a crucial domain of natural language reasoning. In this paper, we evaluate the performance of frontier LLMs in solving physics problems, both mathematical and descriptive. We also employ a plethora of inference-time techniques and agentic frameworks to improve the performance of the models. This includes the verification of proposed solutions in a cumulative fashion by other, smaller LLM agents, and we perform a comparative analysis of the performance that the techniques entail. There are significant improvements when the multi-agent framework is applied to problems that the models initially perform poorly on. Furthermore, we introduce a new evaluation benchmark for physics problems, PHYSICSEVAL, consisting of 19,609 problems sourced from various physics textbooks and their corresponding correct solutions scraped from physics forums and educational websites.

1 Introduction

In the preliminary stages of development, LLMs predominantly exhibited a spuriously correlated reliance on rapid, intuitive processing. As per the dual-process theory by Kahneman (2011)—that posits a fundamental dichotomy between fast, intuitive (System 1) and slower, deliberate (System 2) modes of thought—the inherent System 1 bias of nascent LLMs presented a fundamental weakness in their ability to engage in complex, multistep reasoning. With the advent of prompting techniques such as Chain-of-Thought (CoT) and Tree-of-Thought (ToT), LLMs have bridged this gap and made significant improvements in logical reasoning tasks (Wei et al., 2023; Yao et al., 2023).

Problem:

At what rate does the Sun emit photons? For simplicity, assume that the Sun's entire emission at the rate of 3.9×10^{26} W is at the single wavelength of 550 nm.

Solution:

1. Describe the expression of photon energy.
The energy E of a photon of wavelength λ is given by,

$$E = \frac{hc}{\lambda}$$

Here, h is Planck's constant, and c is the speed of light.

2. Determine the rate of emission of the photon.
Assume that the photons are emitted by a rate R from the sodium lamp. Then, the power P of the sodium lamp is equal to the product of rate R and the energy of each photon E.

$$P = RE \Rightarrow P = R\frac{hc}{\lambda} \Rightarrow R = \frac{P\lambda}{hc} \tag{1}$$

Substitute the below values in Equation 1.
$P = 3.9 \times 10^{26}$ W
$\lambda = 590$ nm
$h = 6.626 \times 10^{-34}$ Js
$c = 3 \times 10^{8}$ m/s
Therefore, the rate of emitted photons from the Sun is:

$$R = \frac{(3.9 \times 10^{26}\ \text{W}) \cdot (550 \times 10^{-9}\ \text{m})}{(6.626 \times 10^{-34}\ \text{Js}) \cdot (3.00 \times 10^{8}\ \text{m/s})}$$
$$= \boxed{1.08 \times 10^{45}\ \text{photons/s}}$$

Figure 1: Example of an astrophysics problem from the PHYSICSEVAL benchmark.

Despite substantial progress in mathematical problem solving, large language models continue to face significant challenges in accurately reasoning through physics problems. Google's generative AI chatbot Bard, based on LaMDA (Thoppilan et al., 2022) and PaLM (Chowdhery et al., 2023) models, ranked in the bottom 2% of test takers on the Physics GRE (Gupta, 2023).

Most research works in recent literature pertaining to improved performance of LLMs in physics include specialized training techniques, such as the method proposed by Anand et al. (2024b). These approaches primarily falter because the LLMs often commit mishaps during their reasoning process, including calculation mistakes, misinterpretation of physical scenarios, and dimensional inconsistencies (Ding et al., 2023a). However, upon adoption of proper prompting heuristics, these mistakes are susceptible to being detected during inference time by LLMs (Pang et al., 2025). This necessitates either a self-correcting framework or an agentic framework with checks and balances that can detect and reconcile the corrigible aspects of a solver model's response. To this end, we utilize a self-refinement technique, where the solver LLM checks its own answer once before generating the final response. In the same vein, we explore different agentic systems where we include a separate, smaller group of LLM agent(s) whose purpose is to review the answer of the solver LLM and provide feedback on probable errors. This method aims to reduce computational overhead while providing an unbiased assessment, addressing the tendency of LLMs to accept their own outputs as correct unless explicitly fine-tuned otherwise (Kadavath et al., 2022). Moreover, this framework enables the use of commercial LLMs for problem-solving while delegating verification to open-source models, thereby minimizing financial overhead via API usage.

In order to facilitate rigorous evaluation of physics reasoning in LLMs, we introduce PHYSICSEVAL, an evaluation benchmark comprising 19,609 physics problems and their elaborated solutions. This dataset is curated by scraping problems and initial solutions from online educational forums, followed by a polishing phase via Google's Gemini 2.5 Pro (Google DeepMind, 2025) to generate detailed, step-by-step explanations. From this collection, we randomly sample 10% ($\approx$ 1,961 problems) to assess the performance of frontier LLMs across four inference-time techniques: baseline inference, self-correction, single-verifier cumulative reasoning, and a multi-agent review framework. The remaining problems are reserved for future model training or fine-tuning efforts.

2 Related Work

The natural language reasoning domain is now confronted with the non-trivial problem, which is that the established mathematical benchmarks are proving to be insufficient. As foundation models like OpenAI's o1 (Jaech et al., 2024) begin to master these evaluations—a challenge anticipated by Phan et al. (2025) who proposed Humanity's Last Exam—it becomes important to distinguish between mere pattern matching and genuine scientific reasoning. Consequently, the field must evolve toward better evaluation benchmarks, not simply to measure capability, but to compel these models to grapple with the kind of difficult problems that actually constitute meaningful progress. The evaluation of LLMs in mathematical reasoning has been characterized by a progressive escalation in benchmark difficulty, driven by advancing model capabilities. Foundational datasets like MAWPS (Koncel-Kedziorski et al., 2016), SVAMP (Patel et al., 2021), ParaMAWPS (Raiyan et al., 2023), GSM8K (Cobbe et al., 2021), and MATH (Hendrycks et al., 2021), which cover up to undergraduate-level content, have been largely surmounted by contemporary models. This performance plateau has spurred the creation of a new tier of challenges. More recent benchmarks, including HARDMath (Fan et al.), OlympiadBench (He et al., 2024), PutnamBench (Tsoukalas et al., 2024), and FrontierMath (Glazer et al., 2024), represent this next frontier, introducing complex university and olympiad-level problems to continue probing the limits of the most advanced systems.

Likewise, it is necessary to move beyond *ad hoc* mathematical problem-solving to evaluate reasoning grounded in the physical world. The domain of physics requires models to integrate an understanding of fundamental laws and real-world principles with the capacity for complex, multistep deductions. Consequently, there is a clear imperative for specialized benchmarks tailored to the unique complexities of advanced physics reasoning. The evolution of physics-reasoning benchmarks for LLMs has progressed systematically from rudimentary problem collections to sophisticated assessment frameworks that evaluate multidimensional aspects of physics comprehension. Initial efforts, exemplified by PhyQA (Ding et al., 2023b) and UGPhysics (Xu et al., 2025), established foundational benchmark suites comprising thousands of structured introductory problems purported to assess basic conceptual understanding. Subsequently, more advanced evaluation instruments such as PhysBench (Qiu et al., 2025) and PhysReason (Zhang et al., 2025) in-

Benchmark	Size	Knowledge	Question		Solution		
			Type	Avg. Tokens	Step-by-step	Avg. Tokens	Avg. Steps
JEEBench	123	CEE	OE, MC	169.7	✗	-	-
MMLU-Pro	1,299	COL	MC	52.1	✓	-	-
GPQA	227	PH.D.	OE	111.4	✗	197.2	3.6
SciEval	1,657	-	OE, MC	154.5	✗	-	-
SciBench	295	COL	OE	80.5	✗	315.9	2.8
MMMU	443	COL	OE, MC	53.8	✗	-	-
ScienceQA	617	K1-K12	MC	13.3	✗	63.0	2.4
OlympiadBench	2,334	COMP	OE	222.0	✗	199.8	3.7
EMMA	156	-	MC	109.5	✗	-	-
PhysReason	1,200	CEE+COMP	OE	226.3	✓	441.3	8.1
UGPhysics	11,040	COL	OE, MC	82.4	✓	318.5	-
PhysicsEval	19,609	CEE+COL+COMP	OE	98.8	✓	3830.8	3.9

Table 1: Comparison of PhysicsEval with various other physics reasoning benchmarks. For the **"Knowledge"** column: COMP refers to Competition level, COL to College level, CEE to College Entrance Examination, K1-K12 to elementary and high school levels, and PH.D. to Doctor of Philosophy. For the **"Question Type"** column: OE denotes open-ended questions, while MC stands for multiple-choice questions.

troduced complex problem formulations requiring extended reasoning chains and multi-step analytical processes. Contemporary benchmark development has culminated in research-oriented assessment suites such as TP-Bench (Chung et al., 2025) and CURIE (Cui et al., 2025), multi-modal frameworks like MMPhyQA (Anand et al., 2024a) that incorporate visual reasoning components, and specialized domain-specific instruments including FEABench (Mudur et al., 2025).

With the intent of eliciting sound reasoning in LLMs for problem-solving, researchers have proposed various methods that try to emulate the thought processes of humans. Cumulative Reasoning (CR) by Zhang et al. (2023), in an iterative fashion involving proposer and verifier LLMs, decomposes problems into smaller and more manageable subproblems and utilizes premises from previous iterations to enhance LLM reasoning. Other works explore self-improvement through verification feedback for better LLM capabilities across diverse domains, including reasoning (Shinn et al., 2023; Hong et al., 2024) and security (Li et al., 2024; Cao et al., 2024). Specifically for physics reasoning, Physics Reasoner (Pang et al., 2025) adopts knowledge-augmentation to facilitate germane formula retrieval, in order to have a properly guided reasoning stage. They also find that the incorporation of checklists acts as a good buttress for LLMs' self-improvement.

3 PhysicsEval Benchmark

To test our inference-time techniques, we curate a new dataset of physics problems, namely PhysicsEval. We compile problems and solutions from various textbooks, spanning from high school to university levels. The list of physics books from which we source the problems is provided in the Appendix C.3. We then use the Gemini 2.5 Pro model by Google[1] to elaborate the solutions into logical steps and sub-steps. The categories of the problems and their key topics are also extracted. Consequently, we are left with a dataset of size 19,609. We partition this dataset into a train set and a test set, maintaining a 90:10 split.

Table 1 depicts a statistical comparison of PhysicsEval with the physics reasoning benchmarks in the existing literature. For the sake of brevity, a complete description of PhysicsEval, including sources, format, examples, comparisons with other datasets, statistics, etc., can be found in Appendix C.

4 Methodology

4.1 Multi-Agent Review Framework

One of the techniques that we use to try to improve the performance of the LLMs is to have several smaller models review their proposed solutions. The subsections delineated below describe this technique in more detail. The method is illustrated in Figure 2.

4.1.1 Proposer Module

Given a set of physics problems $\mathcal{Q} = \{q_1, q_2, \ldots, q_N\}$, the proposer model $\mathcal{P}_\theta$, instantiated as a powerful LLM (such as, Openai's o4-mini or Gemini 2.5 Pro), generates an initial set of solutions $\mathcal{S}^{(0)} = \{s_1^{(0)}, s_2^{(0)}, \ldots, s_N^{(0)}\}$, where each solution is sampled according to

[1] https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/

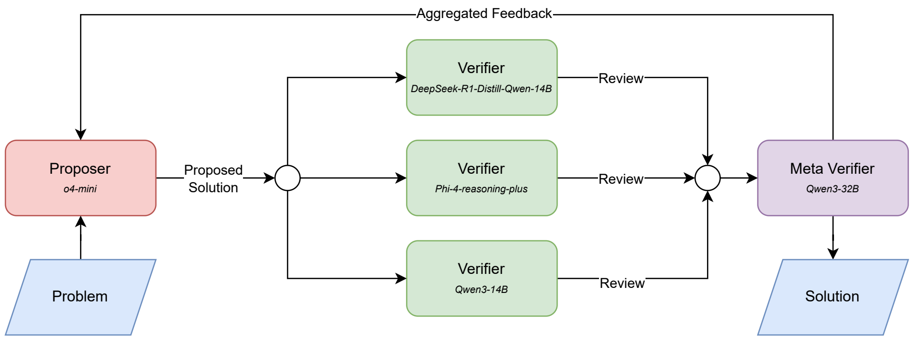

Figure 2: An overview of the multi-agent review model. The model names are, of course, subject to shuffling.

$s_i^{(0)} \sim \mathcal{P}_\theta(\cdot \mid q_i)$. This zero-feedback stage constitutes the base output from which all subsequent verification and refinement occurs.

4.1.2 Verifier Module

To assess the quality of initial solutions generated by the proposer model, we employ a set of three large language models—Microsoft's Phi-4-reasoning-plus (Abdin et al., 2024), Qwen3-14B (Yang et al., 2025), and DeepSeek-R1 14B (Guo et al., 2025) —as independent verifiers. These models operate with fine-tuning and evaluate each proposed solution against the original problem statement.

Each verifier produces a structured evaluation comprising six scores, each ranging from 0 to 5, based on the manually engineered rubrics, inspired by Docktor et al. (2016), outlined in Section 4.4. A weighted average of these scores yields a final scalar score, with greater emphasis placed on formulation (0.25), numerical correctness (0.30), and logical consistency (0.25). The remaining weights are: completeness (0.10), validity of assumptions (0.05), and clarity (0.05). The aggregated score provides a quantitative measure of solution quality, while the individual component scores offer interpretable feedback, which is later utilized by the meta-verifier for cross-verifier analysis and refinement. Each verifier identifies and records perceived mistakes in the proposer's solution, maintaining them in a structured mistake list.

4.1.3 Meta-Verifier Module

The meta-verifier module, Qwen3-32B by Yang et al. (2025), receives the test set questions, proposer solutions, and responses from three independent verifiers. The main task of the meta-verifier is to filter out *irrelevant* or *false* mistakes. Since the verifiers are independent, some mistakes flagged by one verifier may be irrelevant to the actual problem or inaccurate in the context of the solution. The meta-verifier assesses the validity of these mistakes by comparing them across all three verifiers and retains only those that are relevant and consistent. After filtering the mistakes, the meta-verifier aggregates the scores from all three verifiers into a final score using the weighted sum in Equation 2.

$$r_i^{(\text{final})} = 0.5 \times r_i^{(\text{Phi-4})} + 0.3 \times r_i^{(\text{DeepSeek-R1})} + 0.2 \times r_i^{(\text{Qwen3-14B})} \quad (2)$$

This final aggregated score reflects the overall quality of the proposer's solution for each question q_i. The weights were chosen based on the anecdotally observed quality and accuracy of each model's review responses. The meta-verifier's output consists of:

Aggregated Mistakes List: A refined list of mistakes that have been validated across all 3 verifiers.

Aggregated Score: A weighted average of the individual scores from the 3 verifiers, reflecting the overall quality of the proposer's solution. This refined approach helps improve the robustness and accuracy of the evaluation, guiding further refinement of the proposer's solutions based on reliable and consistent feedback.

4.2 Experimental Setup

We use a total of six frontier LLMs to solve 1,962 physics problems from the test split of PHYSICSEVAL. The generated solutions are the base solutions or *proposed* solutions.

4.2.1 Self-refining

After receiving the proposed answer from the solver LLM, we provide its own solution and question again with the following additional metacognitive prompt: "`You are a Physics Professor.`

`Outline the physics principles of the given problem, and please check your own answers for any mistakes, then answer again."` The generated answer is henceforth considered the final answer.

4.2.2 Single-Agent Review

After receiving the proposed answer from the solver LLM, we get another LLM, namely Qwen3-32B, to analyze the question-answer pair and generate a list of probable mistakes. If any mistakes are found, we provide this list as additional information and the solver model's own solution to the solver model itself, and prompt it to solve the problem again.

4.2.3 Multi-Agent Review

We have the reviewer agents generate a list of mistakes for every pair of problem and solution. The meta-verifier agent then compiles these mistakes into a single list. These mistakes are then sent to the solver LLM along with its previous solution, and a new solution is requested, albeit only if the meta-reviewer finds any mistakes. A sample Multi-Agent Review conversation can be found in Appendix B.

4.3 Performance Evaluation and Metrics

For a thorough evaluation of our inference-time techniques and comparison to baseline LLM performance, we use a comprehensive, multi-layered method. This approach breaks down the problem-solving process into key components essential for success in physics, enabling both detailed and overall assessment. Each solution generated by the different LLM configurations (baseline, w/ self-refinement, w/ single-agent verification, and w/ multi-agent verification) is evaluated against the ground-truth solution in PHYSICSEVAL. These ground-truth solutions, derived from established textbook explanations and further elaborated by domain experts for clarity on step-by-step reasoning, serve as the definitive reference for correctness and methodology.

4.4 Rubric Engineering for Evaluation

The evaluation is conducted across six core qualitative and quantitative criteria (similar to the criteria used by the verifier module in multi-agent-verification as shown in Subsection 4.1.2) denoted as M_k where $k \in \{1, \ldots, 6\}$. Each criterion is scored on a Likert scale from 1 to 5 (where 5 represents the highest quality):

Mathematical Accuracy ($S_{\mathbf{MA}}$): Assesses the correctness of calculations, numerical answers, units, and appropriate presentation, strictly in comparison to the ground-truth answer.

Logical Consistency ($S_{\mathbf{LC}}$): Evaluates the soundness of the step-by-step reasoning and its alignment with physics principles and the ground-truth solution's logic.

Completeness ($S_{\mathbf{C}}$): Measures whether all parts of the problem, as scoped and addressed by the ground-truth solution, were fully addressed.

Clarity and Coherence ($S_{\mathbf{CC}}$): Judges the clarity, conciseness, organization, and ease of understanding of the AI's explanation and use of terminology.

Formulas and Principles ($S_{\mathbf{FP}}$): Determines if the correct physical formulas and principles were identified, stated, and applied appropriately by the AI, consistent with the problem's framing in the ground-truth.

Assumptions Made ($S_{\mathbf{A}}$): Assesses whether the *a priori* assumptions were clearly stated, justified, reasonable for the problem context, and did not contradict limitations identified by the ground-truth.

The score for a given solution j on metric k is denoted as $s_{j,k} \in [1, 5]$.

4.5 Physics Proficiency Score (PPS)

To capture overall problem-solving ability, we define a Physics Proficiency Score (PPS)—a weighted average of six key evaluation metrics. For a given solution j, PPS is calculated as:

$$\text{PPS}_j = \sum_{k=1}^{6} w_k \cdot s_{j,k} \qquad (3)$$

where, $s_{j,k}$ denotes the score for solution j on metric k, and w_k is the weight assigned to metric k. The weights reflect the importance of each metric in solving physics problems accurately. These are shown in Table 3. This scoring system emphasizes correctness—especially math, logic, and the proper use of physical principles. Completeness and sound assumptions also matter, while clarity is valued but given less weight. The final PPS is a value that is normalized to be $\in [0, 100]$.

4.6 Evaluation Approach Justification

We use a detailed evaluation framework to closely analyze LLM problem-solving by identifying spe-

	Easy (1-4)				Medium (5-7)				Hard (8-10)			
Model	Baseline	Self-Refine	Single-Agent	Multi-Agent	Baseline	Self-Refine	Single-Agent	Multi-Agent	Baseline	Self-Refine	Single-Agent	Multi-Agent
DeepSeek-R1	90.6	92.7	93.3	**94.1**	80.8	**84.4**	83.1	83.4	72.9	73.7	**74.7**	72.7
Gemma 3 27B	86.9	85.7	86.4	**87.6**	55.8	56.5	**59.4**	59.1	**41.5**	40.4	39.1	40.6
Llama 4 Maverick	91.5	90.5	92.0	**92.9**	**83.6**	82.0	82.6	82.4	55.2	54.0	**57.9**	52.1
Phi-4-reasoning-plus	84.4	94.4	93.2	**94.7**	86.9	92.2	93.4	**93.9**	80.1	83.4	83.8	**87.6**
QwQ-32B	93.7	94.0	94.2	**94.6**	80.9	80.8	81.1	**81.9**	63.9	68.6	63.5	**71.0**
o4-mini	86.7	82.6	85.9	**86.8**	87.3	86.4	87.0	**88.2**	83.6	82.3	84.1	**85.4**
Average	88.96	89.98	90.83	**91.78**	79.21	80.38	81.10	**81.48**	66.20	67.06	67.18	**68.23**

Table 2: Average PPS score on PHYSICSEVAL across frontier LLMs, stratified by problem difficulty and inference method. The "Average" row summarizes results. Higher scores are better. The best-performing method for each model and difficulty tier is highlighted in **bold**.

Metric Component	Weight (w_k)
Mathematical Accuracy (w_{MA})	0.30
Logical Consistency (w_{LC})	0.25
Formulas and Principles (w_{FP})	0.20
Completeness (w_{C})	0.10
Assumptions Made (w_{A})	0.10
Clarity and Coherence (w_{CC})	0.05
Total	**1.00**

Table 3: Weights for Physics Proficiency Score (PPS) sub-metrics

cific strengths and weaknesses. It emphasizes key physics reasoning skills like logic, modeling, and quantitative analysis. Structured rubrics and reference solutions help ensure consistent scoring, balancing objectivity with the subjective nature of some judgments. Weighted scores allow for well-rounded comparisons that align with physics education goals. The framework also supports in-depth error analysis by highlighting reasoning patterns beyond just right or wrong answers.

4.7 Evaluation Process

Solutions were evaluated using Gemini 2.5 Pro. The LLM was provided with a detailed scoring rubric, the ground-truth solution, and the AI-generated solution for each problem. The evaluation prompt provided to Gemini 2.5 Pro was designed to ensure strict comparison against the ground truth. This prompt guided the assignment of scores (1–5 for sub-metrics).

4.8 Statistical Analysis

The collected scores (both for individual metrics $s_{j,k}$ and the aggregated PPS_j) are analyzed using descriptive statistics (mean, median, standard deviation) for each metric and each LLM configuration. Performance is also analyzed across different physics categories and problem complexity levels (proxied by solution length). Table 2 portrays the results yielded by all the pertinent models when evaluated on our PHYSICSEVAL benchmark.

5 Results Analysis

5.1 Key Findings

The evaluation results on the PHYSICSEVAL benchmark, as delineated in Table 2, reveal several key insights into the effectiveness of different agent-based inference methods.

Agent-Based Critique Consistently Improves Performance: A primary observation across all models and difficulty tiers is the general trend of performance improvement with more sophisticated methods. On average, the vanilla Baseline scores are the lowest, followed by incremental gains from Self-Refine, Single-Agent critique, and finally the Multi-Agent framework, which achieves the highest average scores across Easy (91.78), Medium (81.48), and Hard (68.23) problems. This demonstrates that external critique is a more reliable enhancement strategy than simple self-correction.

Gains from Multi-Agent System Amplify with Problem Difficulty: While all methods experience a performance drop as problems become harder, the advantage of the Multi-Agent system over the Baseline becomes more pronounced. For instance, on Hard problems, Phi-4-reasoning-plus gains 7.5 points with the Multi-Agent method over its baseline, and QwQ-32B gains 7.1 points. This suggests that the collaborative verification and feedback process is particularly valuable for tackling complex reasoning tasks where a single model is more likely to fail.

The effectiveness of each method is highly dependent on the underlying model. Phi-4-reasoning-plus stands out as the top-performing model, achieving the highest scores in 7 of the 12 categories. It benefits significantly from all advanced methods, showcasing a strong capacity for both self-correction and assimilating external feedback. o4-mini also proves to be a robust model, establishing the highest baseline for

Comparison	Metric	Weight	t-statistic	p-value	Significant ($p < 0.05$)?
Single Agent vs. Baseline	Overall PPS	1.00	-0.01	9.96e-01	False
Single Agent vs. Baseline	Mathematical Accuracy	0.30	0.89	3.73e-01	False
Single Agent vs. Baseline	Logical Consistency	0.25	0.35	7.29e-01	False
Single Agent vs. Baseline	Completeness	0.10	-2.36	1.82e-02	True
Single Agent vs. Baseline	Clarity And Coherence	0.05	-3.31	9.65e-04	True
Single Agent vs. Baseline	Formulas Principles	0.20	0.29	7.69e-01	False
Single Agent vs. Baseline	Assumptions Made	0.10	-1.23	2.20e-01	False
Multi-Agent Review vs. Baseline	Overall PPS	1.00	2.05	4.05e-02	True
Multi-Agent Review vs. Baseline	Mathematical Accuracy	0.30	2.77	5.70e-03	True
Multi-Agent Review vs. Baseline	Logical Consistency	0.25	1.08	2.78e-01	False
Multi-Agent Review vs. Baseline	Completeness	0.10	-1.44	1.50e-01	False
Multi-Agent Review vs. Baseline	Clarity And Coherence	0.05	-1.79	7.39e-02	False
Multi-Agent Review vs. Baseline	Formulas Principles	0.20	2.50	1.26e-02	True
Multi-Agent Review vs. Baseline	Assumptions Made	0.10	1.50	1.34e-01	False

Table 4: Statistical significance of performance improvements for o4-mini. Results with a p-value < 0.05 are considered statistically significant.

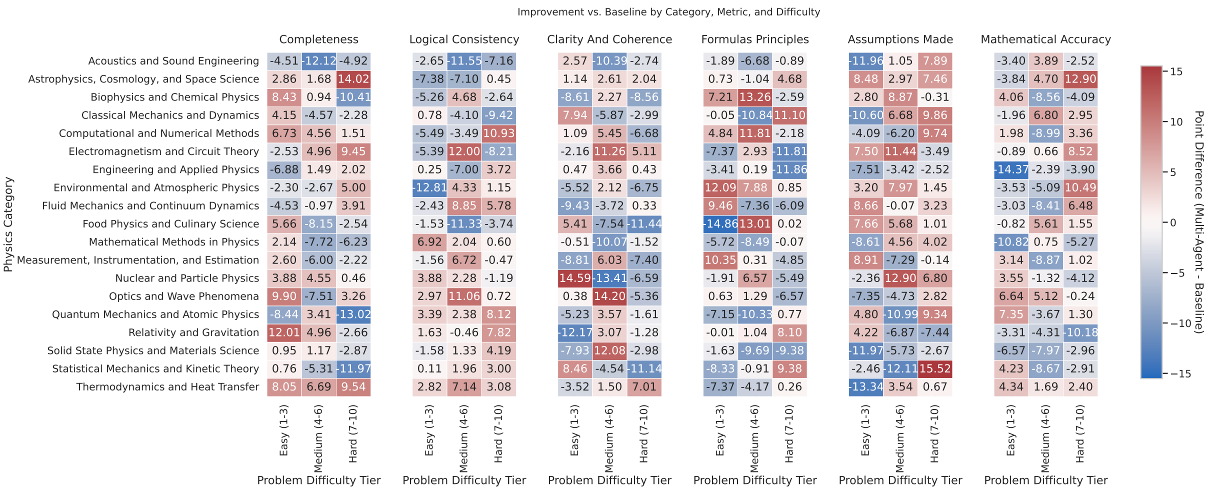

Figure 3: Category-specific impact of the Multi-Agent Review framework across all the scoring rubrics for o4-mini.

Medium and Hard problems and showing consistent improvement with agent-based methods.

A notable and counter-intuitive finding is that the Self-Refine process can harm performance. For models like Gemma 3 27B and Llama 4 Maverick, self-refinement consistently resulted in lower scores compared to their own baseline across all difficulty levels. This indicates that without external guidance, a model's attempt to *"double-check"* its work can introduce new errors or reinforce incorrect initial assumptions, making it an unreliable strategy for certain architectures.

5.2 Results Analysis and Discussion

In order to delve deeper into the performance analysis on PHYSICSEVAL, we keep OpenAI's o4-mini under the limelight. A detailed examination of the model's performance on PHYSICSEVAL reveals a nuanced relationship between the model's inherent capabilities and the efficacy of agent-based refinement methods. The model exhibits a formidable baseline proficiency, with an initial PPS of 85.88, underscored by particularly high scores in Clarity and Coherence (4.76) and Formulas/Principles (4.59). This suggests the model's native strength lies in articulating solutions clearly and correctly identifying the underlying physics. However, the Self-Refined approach proved detrimental, leading to a notable decline in the overall PPS to 84.58. A closer look at the sub-metrics uncovers that while Mathematical Accuracy saw a slight improvement (4.17 to 4.22), this came at a significant cost to Completeness (4.56 to 4.40) and Clarity (4.76 to 4.54). This paradox indicates that without external guidance, the model's attempt to self-correct can disrupt the solution's logical structure and narrative coherence. In contrast, while the Single Agent method offered negligible improvement, the Multi-Agent Review emerges as the only method to yield a definitive performance enhancement, elevating the PPS to 86.84. This gain is not merely marginal but is driven by targeted improvements in the model's weakest areas—Mathematical Accuracy (4.24), Logical Consistency (4.56), and Assumptions Made (4.43)—while largely preserving its strong baseline clarity. This finding powerfully suggests that for a highly capable model like o4-mini, further advancement is not achieved through simple self-correction but through a robust, consensus-driven verification process that can surgically address specific logical and computational flaws without compromising the solution's overall quality.

5.2.1 Category-Specific Impact of the Multi-Agent Framework

The impact of Multi-Agent Review on o4-mini across different categories can be seen in Figure 3. It reveals that the impact of multi-agent review on o4-mini's performance is highly category- and rubric-dependent, with both substantial gains and notable degradations across physics domains and difficulty tiers. Noteworthy improvements are observed in categories such as Quantum Mechanics and Atomic Physics, Relativity and Gravitation, and Thermodynamics and Heat Transfer, particularly for hard problems and rubrics like Completeness and Formulas/Principles. Medium-difficulty problems in areas like Food Physics and Culinary Science and Optics, and Wave Phenomena also see dramatic gains in Clarity and Coherence. However, the benefits are not universal: some categories, including Acoustics and Sound Engineering and Engineering and Applied Physics, experience negative or inconsistent changes, especially for easier problems, and certain rubrics, such as Assumptions Made and Mathematical Accuracy, even show degradations in select domains. These results highlight that while multi-agent review can substantially enhance performance in specific contexts—especially for complex, technical, or computationally intensive tasks—it may also introduce confusion or diminish quality in others, underscoring the need for more adaptive and context-aware review strategies. Additional details can be seen in Figure 4 in Appendix C.4.

5.2.2 Statistical Significance of Performance Changes

To evaluate the effectiveness of different inference-time techniques, we perform paired t-tests comparing each method against the o4-mini baseline, using a significance threshold of $p < 0.05$. The Single Agent method fails to produce any statistically significant improvement in Overall PPS ($p = 0.996$), and instead shows significant degradations in Completeness ($p = 0.0182$) and Clarity and Coherence ($p = 0.000965$). This suggests that introducing a single external reviewer may negatively impact the structural and narrative quality of the generated solutions, potentially by injecting inconsistent or insufficient feedback. On the other hand, the Multi-Agent Review approach demonstrated statistically significant improvements in Overall PPS ($p = 0.0405$), Mathematical Accuracy ($p = 0.0057$), and Formulas Principles ($p = 0.0126$). These gains indicate that collaborative critique across multiple agents can more effectively guide the model toward better mathematical correctness and principled reasoning. Although improvements in other dimensions, such as Logical Consistency and Completeness, do not reach significance, the overall results imply the advantage of multi-agent systems in enhancing both the accuracy and interpretability of the model's solutions.

6 Conclusion

Our comprehensive evaluation of frontier LLMs on the newly introduced PhysicsEval benchmark demonstrates both the promise and the current limitations of LLMs in the domain of physics problem-solving. While baseline model performance is already strong for many categories, our experiments reveal that agentic inference-time techniques—particularly multi-agent verification—can yield substantial improvements, especially for challenging problems and in technical subfields where initial model outputs are less reliable. However, these gains are not uniform across all categories and rubrics, highlighting the nuanced and context-dependent nature of collaborative critique. Our findings underscore the importance of adaptive, category-aware strategies for further advancing LLM capabilities in scientific reasoning. We hope that PhysicsEval will serve as a valuable resource for the research community.

Limitations

One main limitation of our approach is the high computational cost. The multi-agent review method runs several models for each question, which increases processing time and resource use. This makes it less practical for real-time or low-resource settings. Moreover, our dataset, while large and varied, didn't undergo full manual checking. The solutions are expanded using an LLM (Gemini 2.5 Pro), and only a small sample is reviewed. This means there could be mistakes in the data that affect model performance and evaluation. Due to the elaborate and descriptive nature of the ground truth, the evaluation of the solutions is largely LLM based. Lastly, while our methods work well for physics problems, they may not transfer easily to other STEM areas without changes or fine-tuning.

Ethics Statement

This work involves the use of publicly available large language models and does not include any human subjects, private data, or personally identifiable information. All physics problems are collected from publicly available sources, and care is taken to ensure that no copyrighted or proprietary content is used without proper attribution. The dataset partially relies on LLM-generated content for elaboration, which may introduce unintended biases or inaccuracies. We acknowledge this limitation and emphasize the importance of responsible use and validation in downstream applications. The proposed multi-agent inference techniques aim to improve performance without retraining, allowing broader access to high-performing systems. However, we recognize the increased computational cost associated with such approaches and encourage mindful deployment, especially in energy-sensitive or resource-constrained environments. We support transparency and reproducibility and plan to release the dataset and code where licensing permits.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Avinash Anand, Janak Kapuriya, Apoorv Singh, Jay Saraf, Naman Lal, Astha Verma, Rushali Gupta, and Rajiv Ratn Shah. 2024a. Mm-phyqa: Multimodal physics question-answering with multi-image cot prompting. In *Advances in Knowledge Discovery and Data Mining: PAKDD 2024, Taipei, Taiwan, May 7–10, 2024, Proceedings, Part V*, volume 14649 of *Lecture Notes in Computer Science*, pages 53–64, Singapore. Springer Singapore.

Avinash Anand, Kritarth Prasad, Chhavi Kirtani, Ashwin R Nair, Mohit Gupta, Saloni Garg, Anurag Gautam, Snehal Buldeo, and Rajiv Ratn Shah. 2024b. Enhancing llms for physics problem-solving using reinforcement learning with human-ai feedback.

Haim Baruh. 2015. *Applied dynamics*. CRC press Boca Raton.

Richard E Berg and David G Stork. 2012. *The physics of sound*. Pearson Education India.

Mary L Boas. 2006. *Mathematical methods in the physical sciences*. John Wiley & Sons.

William D Callister and David G Rethwisch. 2022. *Fundamentals of materials science and engineering*. John Wiley & Sons.

Chentao Cao, Zhun Zhong, Zhanke Zhou, Yang Liu, Tongliang Liu, and Bo Han. 2024. Envisioning outlier exposure by large language models for out-of-distribution detection. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 5629–5659. PMLR.

Sean M Carroll. 2004. An introduction to general relativity: spacetime and geometry. *Addison Wesley*, 101:102.

Yunus A Cengel and Afshin J Ghajar. 2014. *Heat and Mass Transfer (in SI units)*. Mcgraw-Hill Education-Europe, London.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Daniel JH Chung, Zhiqi Gao, Yurii Kvasiuk, Tianyi Li, Moritz Münchmeyer, Maja Rudolph, Frederic Sala, and Sai Chaitanya Tadepalli. 2025. Theoretical physics benchmark (tpbench)–a dataset and study of ai reasoning capabilities in theoretical physics. *arXiv preprint arXiv:2502.15815*.

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

Hao Cui, Zahra Shamsi, Gowoon Cheon, Xuejian Ma, Shutong Li, Maria Tikhanovskaya, Peter Christian Nørgaard, Nayantara Mudur, Martyna Beata Plomecka, Paul Raccuglia, Yasaman Bahri, Victor V. Albert, Pranesh Srinivasan, Haining Pan, Philippe Faist, Brian Rohr, Michael J. Statt, Dan Morris, Drew Purves, Elise Kleeman, Ruth Alcantara, Matthew Abraham, Muqthar Mohammad, Ean Phing VanLee, Chenfei Jiang, Elizabeth Dorfman, Eun-Ah Kim, Michael P. Brenner, Viren Jain, Sameera Ponda, and Subhashini Venugopalan. 2025. Curie: Evaluating llms on multitask scientific long-context understanding and reasoning. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, page –, Addis Ababa, Ethiopia. Accepted to ICLR 2025.

Jingzhe Ding, Yan Cen, and Xinyuan Wei. 2023a. Using large language model to solve and explain physics word problems approaching human level.

Jingzhe Ding, Yan Cen, and Xinyuan Wei. 2023b. Using large language model to solve and explain physics word problems approaching human level. *arXiv preprint arXiv:2309.08182*.

Jennifer L Docktor, Jay Dornfeld, Evan Frodermann, Kenneth Heller, Leonardo Hsu, Koblar Alan Jackson, Andrew Mason, Qing X Ryan, and Jie Yang. 2016. Assessing student written problem solutions: A problem-solving rubric with application to introductory physics. *Physical review physics education research*, 12(1):010130.

Leah Edelstein-Keshet. 2005. *Mathematical models in biology*. SIAM.

Jingxuan Fan, Sarah Martinson, Erik Y Wang, Kaylie Hausknecht, Jonah Brenner, Danxian Liu, Nianli Peng, Corey Wang, and Michael Brenner. Hardmath: A benchmark dataset for challenging problems in applied mathematics. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Douglas C Giancoli. 2005. *Physics: principles with applications*, volume 1. Pearson Educación.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. 2024. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*.

Google DeepMind. 2025. Gemini 2.5 pro: Our most intelligent ai model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. Accessed: 2025-07-27.

David J Griffiths and Darrell F Schroeter. 2018. *Introduction to quantum mechanics*. Cambridge university press.

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

Pranav Gupta. 2023. Testing llm performance on the physics gre: some observations.

David Halliday, Robert Resnick, and Jearl Walker. 2013. *Fundamentals of physics*. John Wiley & Sons.

William H Hayt Jr and John A Buck. 2001. *Engineering Electromagnetics Sixth Edition William H. Hayt, Jr.. John A. Buck, 2001 The McGraw Companies*. The McGraw Companies.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Ruixin Hong, Hongming Zhang, Xinyu Pang, Dong Yu, and Changshui Zhang. 2024. A closer look at the self-verification abilities of large language models in logical reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 900–925, Mexico City, Mexico. Association for Computational Linguistics.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, NY, USA.

Laura Kay, Stacy Palen, and George Blumenthal. 2016. *21st century astronomy*. WW Norton & Company.

Randall Dewey Knight. 2015. *Physics for scientists and engineers*. W. Ross MacDonald School Resource Services Library.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. Mawps: A math word problem repository. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 1152–1157.

Kenneth S Krane. 2019. *Modern physics*. John Wiley & Sons.

Xuan Li, Zhanke Zhou, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. 2024. Deepinception: Hypnotize large language model to be jailbreaker. In *Neurips Safe Generative AI Workshop 2024*.

Nayantara Mudur, Hao Cui, Subhashini Venugopalan, Paul Raccuglia, Michael P Brenner, and Peter Norgaard. 2025. Feabench: Evaluating language models on multiphysics reasoning ability. *arXiv preprint arXiv:2504.06260*.

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, Abu Dhabi, UAE. Association for Computational Linguistics.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

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

Syed Rifat Raiyan, Md Nafis Faiyaz, Shah Md. Jawad Kabir, Mohsinul Kabir, Hasan Mahmud, and Md Kamrul Hasan. 2023. Math word problem solving by generating linguistic variants of problem statements. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 362–378, Toronto, Canada. Association for Computational Linguistics.

Robert Resnick. 1989. *Introduction to Special Relativity*, 1st edition. John Wiley & Sons, New York.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36: Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, volume 36 of *Advances in Neural Information Processing Systems*, pages 8634–8652, New Orleans, LA, USA. Neural Information Processing Systems Foundation.

John R. Taylor. 2006. *Classical Mechanics: An Undergraduate Text*, 1st edition. University Science Books, Sausalito, CA.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings, Amitayush Thakur, and Swarat Chaudhuri. 2024. Putnambench: A multilingual competition-mathematics benchmark for formal theorem-proving. In *AI for Math Workshop@ ICML 2024*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Frank M White. 2012. Fluid mechanics, in si units.

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. *arXiv preprint arXiv:2502.00334*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Hugh D Young, Roger A Freedman, and Ragbir Bhathal. 2010. *University physics: Australian edition*. Pearson Higher Education AU.

Nouredine Zettili. 2009. Quantum mechanics: concepts and applications.

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. *arXiv preprint arXiv:2502.12054*.

Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. 2023. Cumulative reasoning with large language models. *arXiv preprint arXiv:2308.04371*.

A Evaluation Prompt for Gemini 2.5 Pro

You are an expert physics problem evaluator. Your task is to meticulously and STRICTLY compare an AI-generated solution to a manual, ground-truth solution for a given physics problem. The Ground Truth Solution is considered the definitive correct answer and approach for the given problem statement. Deviations by the AI-Generated Solution from the Ground Truth, especially in terms of method, assumptions, interpretation of given data, or parts deemed unsolvable by the Ground Truth, MUST be penalized appropriately according to the guidelines below.

Evaluate the AI-generated solution based on the following categories and scoring guidelines. Provide your evaluation STRICTLY as a JSON object.

Evaluation Categories and Scoring Guidelines:

1. mathematical_accuracy: (Score 1-5) How correct are the AI's calculations, numerical answers, and units *when compared to the problem defined by the Ground Truth*?

- 5: All calculations, numerical results, and units are perfectly correct and appropriately presented, AND align with the Ground Truth's final answers if the same method is used, OR are verifiably correct if a different valid method is used.

- 4: Minor calculation error in the AI solution, or an incorrect/missing unit, but the AI's underlying mathematical method (if aligned with GT or validly alternative) is sound.

- 3: Several minor errors in the AI solution, or one significant calculation error that impacts the AI's result. Units might be inconsistently handled.

- 2: Major calculation errors or fundamental misunderstandings of mathematical operations in the AI solution. If the AI solution uses different input data values than implied by the Ground Truth (e.g., different length, mass), leading to numerically different answers, score 2 here even if its internal math is correct for its chosen data, because it's not solving the *Ground Truth's* problem.

- 1: Almost all calculations in the AI solution are incorrect, non-sensical, or missing. The AI uses drastically different input data leading to completely irrelevant numerical results for the Ground Truth problem.

2. logical_consistency: (Score 1-5) Does the AI solution follow a logical step-by-step progression? Is the AI's reasoning sound and aligned with physics principles, *ideally mirroring or compatibly extending the Ground Truth's logic*?

- 5: The AI solution flows perfectly. Each step logically follows from the previous one. The reasoning is impeccable and aligns well with the Ground Truth's approach or a valid alternative.

- 4: AI solution is mostly logical and well-reasoned. Perhaps one step is slightly unclear or its justification is weak, but it doesn't break the overall logic or significantly deviate from a valid path.

- 3: Some logical gaps, inconsistencies, or steps in the AI solution that don't clearly follow, making the solution harder to follow or verify, or deviating from the core logic of the Ground Truth without clear justification.

- 2: Significant logical flaws in the AI solution. Steps are out of order, reasoning is poor or contradictory to established physics or the Ground Truth's interpretation.

- 1: The AI solution is illogical, incoherent, or internally contradictory.

3. completeness: (Score 1-5) Does the AI-generated solution address all parts of the problem *as understood and scoped by the Ground Truth*?

- 5: All parts of the problem (including sub-questions, if any), as addressed or implied as solvable by the Ground Truth, are fully addressed and answered by the AI.

- 4: A minor aspect of the problem (as per GT) is overlooked by the AI, or one sub-question is not fully answered or is missing.

- 3: A significant part of the problem (as per GT) is ignored or left unanswered by the AI. If the Ground Truth indicates a part of the problem is unsolvable with given data, but the AI attempts to solve it by making significant unstated/unwarranted assumptions, this is a flaw in understanding problem scope; score 3 or lower.

- 2: Only a small portion of the problem (as per GT) is addressed by the AI; major components are missing.

- 1: The problem is largely unaddressed by the AI, or the AI solution is off-topic relative to the Ground Truth.

4. clarity_and_coherence: (Score 1-5) Is the AI's explanation clear, concise, and easy to understand?

- 5: The AI explanation is exceptionally clear, concise, well-structured, and very easy to understand. Excellent use of language and terminology.

- 4: The AI explanation is clear and generally easy to understand, with minor areas for improvement in conciseness, structure, or flow.

- 3: The AI explanation is generally understandable but may be verbose, unclear in parts, poorly organized, or contain jargon without adequate explanation.

- 2: The AI explanation is difficult to understand due to ambiguity, poor writing, or convoluted structure.

- 1: The AI explanation is incomprehensible, extremely poorly written, or nonsensical.

5. formulas_principles: (Score 1-5) Are correct physical formulas and principles identified and applied correctly by the AI, *and are they appropriate for the problem as framed by the Ground Truth*?

- 5: All necessary physical formulas and principles are correctly identified, stated, and applied appropriately by the AI, consistent with the Ground Truth's approach or a valid, equally rigorous alternative.

- 4: Mostly correct formulas/principles used by AI. Perhaps a minor error in recalling a formula, or a slight misapplication of a correct principle that doesn't fundamentally alter the solution path compared to GT.

- 3: Some incorrect formulas/principles are used by AI, or correct ones are applied incorrectly in a significant way. Or, the AI uses a principle that oversimplifies the problem compared to the level of detail expected by the Ground Truth.

- 2: Major errors in formula/principle selection or application by AI. Fundamental physics concepts are misunderstood by the AI.

- 1: Completely inappropriate formulas/principles are used by AI, or relevant physics is entirely ignored.

6. assumptions_made: (Score 1-5) Are AI assumptions (explicit or implicit) explicit, justified, and reasonable *especially when compared to the Ground Truth's scope and stated/implied assumptions*?

- 5: All necessary assumptions made by the AI are explicitly stated, well-justified, and perfectly reasonable for the problem context, AND do not contradict or bypass limitations identified by the Ground Truth.

- 4: Most necessary assumptions made by the AI are stated and reasonable; some minor ones might be implicit but obvious, or lack full justification but are acceptable and align with GT.

- 3: Some key assumptions in the AI solution are missing, not clearly stated, or questionable in reasonableness. Or, the AI makes assumptions that simplify the problem in a way the Ground Truth does not.

- 2: Major unreasonable assumptions are made by the AI, or critical assumptions are not stated, leading to an incorrect or flawed solution path. This includes assumptions that allow solving parts the Ground Truth indicates are unsolvable with the given data.

- 1: Assumptions in the AI solution are entirely inappropriate, absent when clearly needed, or lead to a trivialization/misrepresentation of the problem as defined by the Ground Truth.

7. overall_correctness: (Score 0-10) How correct and sound is the AI's approach and final answer(s) overall, *primarily judged by its fidelity to the Ground Truth's interpretation, method, and result for the specific problem*?

- 10: Perfect solution. The AI's method, reasoning, data interpretation, assumptions, and final answer(s) align flawlessly or are an equally valid and rigorous path to the Ground Truth.

- 8-9: Excellent solution. Fundamentally correct with very minor, inconsequential flaws or slight stylistic deviations from the Ground Truth, but arrives at the same essential understanding and results.

- 6-7: Good solution. Generally correct approach by the AI, and largely correct answer(s), but with some noticeable errors, omissions, or areas for improvement when compared to the Ground Truth. The AI might use a valid but less ideal method.

- 4-5: Partially correct. The AI demonstrates some understanding but contains significant flaws in reasoning, calculation, choice of principles, or makes unwarranted assumptions that lead it away from the Ground Truth's solution. This score is appropriate if the AI solves a simplified version of the problem or misses key constraints implied by the Ground Truth.

- 2-3: Mostly incorrect. The AI shows fundamental misunderstandings of the problem or physics principles as defined by the Ground Truth. A solution that uses

different fundamental input data than the Ground Truth CANNOT be rated higher than 3, even if its internal logic is sound for its chosen data.

- 0-1: Completely incorrect, irrelevant, or no meaningful attempt made by the AI to solve the problem as presented and solved by the Ground Truth.

Problem ID: problem_id

Ground Truth Solution (this is the reference correct solution):

ground_truth

Elaborated Solution Steps (this explains the Ground Truth):

Elaborated Solution Steps (Manual): elaborated_solution

AI-Generated Solution to Evaluate (compare this against the Ground Truth and Elaborated Solution): ai_solution

Provide your evaluation STRICTLY as a JSON object with the problem_id and scores for each category listed above. Your entire response should be ONLY the JSON object, starting with and ending with .

```
Example JSON format:
{{
    "problem_id": "{problem_id}",
    "mathematical_accuracy": <
        score_1_to_5>,
    "logical_consistency": <score_1_to_5
        >,
    "completeness": <score_1_to_5>,
    "clarity_and_coherence": <
        score_1_to_5>,
    "formulas_principles": <score_1_to_5
        >,
    "assumptions_made": <score_1_to_5>,
    "overall_correctness": <
        score_0_to_10>
}}
```

B Sample Conversation with o4-mini (Multi-Agent Review)

Here, we have done multi agent review on a proposed solution. The list of mistakes generated by the review process is fed back to the proposer.

User: You are an expert on Physics. You solve problems step by step while maintaining logical consistency. Solve the following Physics problem: A toroidal inductor with an inductance of 9.0 mH encloses a volume of 0.0200 m^3. If the average energy density in the toroid is 70.0 J/m^3, what is the current through the inductor? Finally, write the final answers in brief. Make sure you write all equations in LaTeX.

Assistant (o4-mini - Initial Solution): Here is the step-by-step solution: 1. Energy stored in an inductor $U = \frac{1}{2}LI^2$ 2. Relation between total energy and energy density $U = uV$ where u = energy density, V = volume 3. Equate the two expressions for U: $\frac{1}{2}LI^2 = uV$ 4. Solve for the current I: $I = \sqrt{\frac{2uV}{L}}$ 5. Substitute the given values: $L = 9.0 \times 10^{-3}$ H $u = 70.0$ J/m^3 $V = 0.0200$ m^3 $I = \sqrt{\frac{2 \cdot 70.0 \cdot 0.0200}{9.0 \times 10^{-3}}} = \sqrt{\frac{2.80}{9.0 \times 10^{-3}}} = \sqrt{311.11} \approx$ 17.6 A

Final Answer (brief): $I = \sqrt{\frac{2uV}{L}} \approx 17.6$ A

User (Feedback from Meta-Verifier based on Reviewers): I have some feedback. The solution assumes that the average energy density is uniformly distributed throughout the entire volume of the toroid. In reality, for a toroid, the magnetic field varies inversely with the distance from the center, leading to non-uniform energy density. The formula used assumes uniform energy density, which is an approximation. This assumption might not be accurate if the field varies significantly within the toroid. After taking this into account, please generate the solution once again. Remember to write all equations in LaTeX.

[o4-mini would then generate a revised solution, ideally addressing the feedback.]

C PhysicsEval : Additional Details

C.1 Construction

To enable large-scale evaluation and training of reasoning-capable language models in physics, we curate a comprehensive dataset of **19,609 annotated problems**, sourced from 20 different authoritative physics textbooks and verified educational websites.

The dataset spans 19 different categories, including *Mechanics, Thermodynamics, Electromagnetism, Waves, Optics, Relativity*, and *Quantum Physics.*

Each problem is processed through the following pipeline:

- **Data Cleaning:** Raw content is cleaned to remove noise and inconsistencies.
- **LaTeXAnnotation:** All equations are converted into LaTeXfor structured mathematical representation.
- **Step-Wise Elaboration:** Using Gemini 2.5 Pro in "Think" mode, solutions are decomposed into logically coherent steps to enhance interpretability for LLMs.
- **Metadata Tagging:** Each problem is annotated with topic category, difficulty level, and key physical principles.

Train-Test Split: We apply a 90:10 split, resulting in **17,647 training** and **1,962 test samples**, supporting generalization across diverse reasoning tasks.

The problems in the dataset are given a difficulty score from 1 to 10. The number of steps in the elaborated solution is stored as well. Some alternative solution methods are also suggested.

C.2 Data Model

The dataset has the following fields:

- **Problem_ID**: Unique identifier for the problem instance
- **problem**: Original, full problem text from source material
- **simplified_problem_statement**: Paraphrased version, stripped of complexity
- **category**: Topical category (*e.g.*, Mechanics, Optics)
- **soft_labels**: Tags like numerical, conceptual, multi-step, diagram
- **elaborated_solution_steps**: Step-by-step reasoning to the correct answer
- **alternative_solutions**: Different valid solution methods
- **problem_difficulty**: Difficulty rating (1–10)
- **final_answers_in_brief**: Final answer(s) only, no reasoning
- **steps**: Number of steps in main solution
- **source**: The source of the problem

C.3 Dataset Sources

- **21st Century Astronomy: Stars and Galaxies, 4th Edition** - by Laura Kay, Stacy Palen, Brad Smith, and George Blumenthal (Kay et al., 2016)
- **A Complete Resource Book for JEE Main 2018: Physics** - by Sanjeev Kumar
- **Physics: Principles with Applications, 7th Edition** - by Douglas C. Giancoli (Giancoli, 2005)
- **Physics for Scientists and Engineers: A Strategic Approach with Modern Physics, 4th Edition** - by Randall D. Knight (Knight, 2015)
- **Mathematical Models in Biology, 1st Edition** - by Leah Edelstein-Keshet (Edelstein-Keshet, 2005)
- **Fundamentals of Physics, 10th Edition** - by David Halliday, Robert Resnick, and Jearl Walker (Halliday et al., 2013)
- **Mathematical Methods in the Physical Sciences, 3rd Edition** - by Mary L. Boas (Boas, 2006)
- **Heat and Mass Transfer: Fundamentals and Applications, 5th Edition** - by Yunus A. Çengel and Afshin J. Ghajar (Cengel and Ghajar, 2014)
- **Materials Science and Engineering: An Introduction, 8th Edition** - by William D. Callister Jr. (Callister and Rethwisch, 2022)
- **Fluid Mechanics in SI Units, 8th Edition (2017)** - by Frank M. White (White, 2012)
- **University Physics with Modern Physics, 14th Edition** - by Hugh D. Young and Roger A. Freedman (Young et al., 2010)
- **Analytical Dynamics, 1st Edition** - by Haim Baruh (Baruh, 2015)
- **Engineering Electromagnetics, 8th Edition** - by William H. Hayt Jr. and John A. Buck (Hayt Jr and Buck, 2001)
- **Modern Physics, 2nd Edition** - by Kenneth S. Krane (Krane, 2019)
- **Introduction to Quantum Mechanics, 2nd Edition** - by David J. Griffiths (Griffiths and Schroeter, 2018)
- **Quantum Mechanics: Concepts and Applications, 2nd Edition** - by Nouredine Zettili (Zettili, 2009)
- **Classical Mechanics: An Undergraduate Text, 2006 Edition** - by John R. Taylor (Taylor, 2006)
- **Introduction to Special Relativity, 1989 Edition** - by Robert Resnick (Resnick, 1989)
- **The Physics of Sound, 3rd Edition** - by Richard E. Berg and David G. Stork (Berg and Stork, 2012)
- **Spacetime and Geometry: An Introduction to General Relativity, 2004 Edition** - by Sean M. Carroll (Carroll, 2004)

C.4 Further Information

Field Name	Strength / Purpose
`Problem_ID`	Enables consistent referencing, indexing, and analysis across models and experiments.
`problem`	Maintains fidelity to real-world phrasing typically encountered in education or exams.
`simplified_problem_statement`	Helps models/annotators focus on core reasoning, improving interpretability.
`category`	Enables domain-wise evaluation and curriculum design.
`soft_labels`	Supports nuanced supervision and better error analysis.
`elaborated_solution_steps`	Crucial for explainability and multi-step reasoning evaluation.
`alternative_solutions`	Promotes robustness and exposure to diverse reasoning.
`problem_difficulty`	Enables benchmarking and curriculum learning by difficulty.
`final_answers_in_brief`	Useful for accuracy checks and extractive answer training.
`steps`	Helps with curriculum learning and complexity analysis.

Table 5: Description of fields in the physics dataset and their purposes

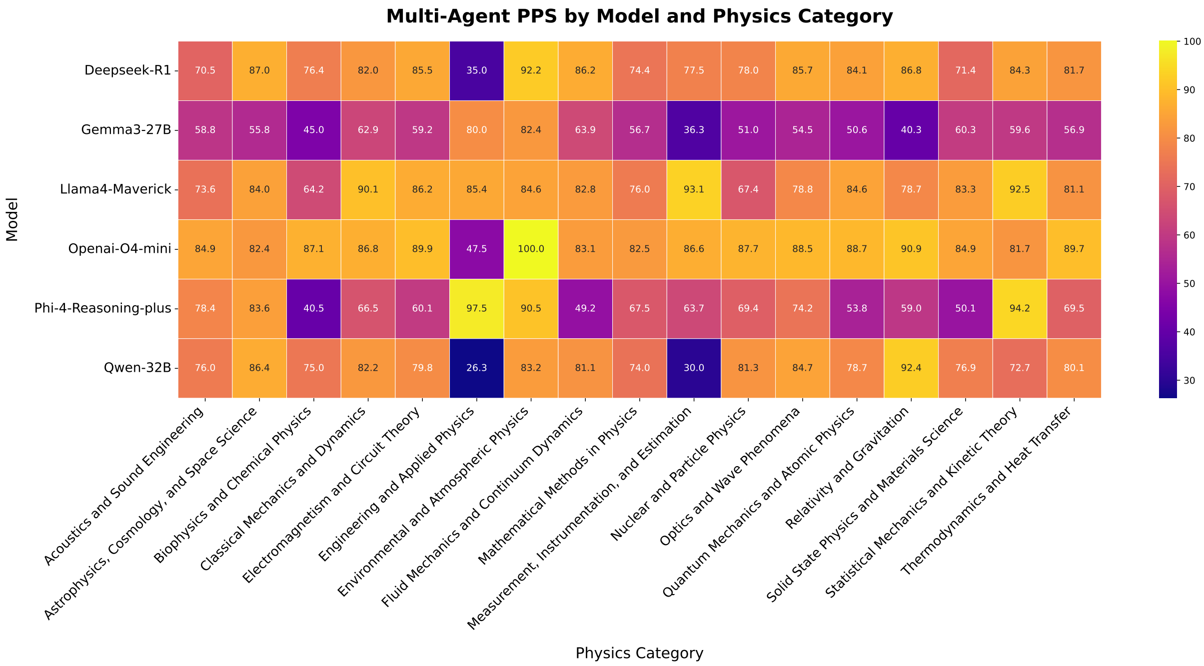

Figure 4: Multi-Agent PPS by Model and Physics Category

Statistic	problem_difficulty	steps	problem_tokens	solution_tokens
Count	19609	19609	19609	19609
Mean	5.720282	3.883523	98.815595	3830.767403
Std Dev	1.445578	1.776616	78.899039	2458.892157
Min	1.000000	0.000000	1.000000	311.000000
25%	5.000000	3.000000	51.000000	2422.000000
50% (Median)	6.000000	4.000000	81.000000	3115.000000
75%	7.000000	5.000000	124.000000	4300.000000
Max	10.000000	23.000000	4380.000000	29931.000000

Table 6: Descriptive statistics for problem dataset

Problem Category	Total Problems	Test Set
Acoustics and Sound Engineering	589	58
Quantum Mechanics and Atomic Physics	1677	155
Thermodynamics and Heat Transfer	2451	234
Solid State Physics and Materials Science	789	78
Fluid Mechanics and Continuum Dynamics	936	95
Electromagnetism and Circuit Theory	2791	259
Optics and Wave Phenomena	1301	132
Classical Mechanics and Dynamics	4103	443
Nuclear and Particle Physics	766	76
Statistical Mechanics and Kinetic Theory	171	18
Astrophysics, Cosmology, and Space Science	961	105
Relativity and Gravitation	656	68
Mathematical Methods in Physics	1991	193
Biophysics and Chemical Physics	148	18
Environmental and Atmospheric Physics	43	5
Measurement, Instrumentation, and Estimation	171	17
Engineering and Applied Physics	41	5
Computational and Numerical Methods	13	1
Food Physics and Culinary Science	11	2

Table 7: Categories of problems and their amounts in the dataset

Example from the Dataset

Problem_ID

46b6dfac-1f0c-4e23-9230-798ce854e963

problem

A long homogeneous resistance wire of radius $r_o = 5\ \mathrm{mm}$ is being used to heat the air in a room by the passage of electric current. Heat is generated in the wire uniformly at a rate of $5 \times 10^7\ \mathrm{W/m^3}$ ... ... Take the thermal conductivity of the wire to be $k = 6\ \mathrm{W/m \cdot K}$.

simplified_problem_statement

A long wire with a radius of 5 mm generates heat uniformly at a rate of 5 x $10\hat{7}$ W/$m\hat{3}$. The outer surface temperature of the wire is maintained at 180°C. Determine the temperature at a distance of 3.5 mm from the center of the wire, given that the thermal conductivity of the wire is 6 W/m·K.

category

Thermodynamics and Heat Transfer

soft_labels

- Heat Generation
- Steady-State Heat Transfer
- Radial Heat Conduction
- Boundary Conditions
- Thermal Conductivity

elaborated_solution_steps

Certainly! As a Professor of Physics, I'd be glad to elaborate on each step of the provided solution to ensure you understand the underlying physics and mathematical operations thoroughly. Let's break down each step:

**Step 01: The heat generat... ...owever needs correction for accurate temperature distribution. But as instructed, I have elaborated on each step as provided in the solution, highlighting the physics and math, without altering the given equations.

alternative_solutions

- One could use a finite element method to solve the heat equation numerically, especially if the geometry or boundary conditions were more complex.

problem_difficulty

6

final_answers_in_brief

- Temperature at r=3.5 mm is 178.86 °C

steps

6

source

Heat and Mass Transfer: Fundamentals and Applications, 5th Edition - by Yunus A. Çengel and Afshin J. Ghajar

problem_tokens

171

solution_tokens

11736