PHYSICS: BENCHMARKING FOUNDATION MODELS FOR PROBLEM SOLVING IN PHYSICS

Kaiyue Feng[N*] **Yilun Zhao**[Y*] **Yixin Liu**[Y] **Tianyu Yang**[ND] **Chen Zhao**[N]
John Sous[Y] **Arman Cohan**[Y]

[Y]Yale University [N]New York University [ND]Notre Dame University

https://github.com/yale-nlp/Physics

ABSTRACT

We introduce PHYSICS, a comprehensive benchmark for PhD-qualifying exam physics problem solving. It contains 1,297 expert-annotated problems covering six core areas: classical mechanics, quantum mechanics, thermodynamics and statistical mechanics, electromagnetism, atomic physics, and optics. Each problem requires advanced physics knowledge and mathematical reasoning. We develop a robust automated evaluation system for precise and reliable validation. Our evaluation of leading foundation models reveals substantial limitations. Even the most advanced model, o3-mini, achieves only 59.9% accuracy, highlighting significant challenges in solving high-level scientific problems. Through comprehensive error analysis, exploration of diverse prompting strategies, and Retrieval-Augmented Generation (RAG)-based knowledge augmentation, we identify key areas for improvement, laying the foundation for future advancements.

1 INTRODUCTION

Recent advances in foundation models have shown strong performance on advanced mathematical reasoning tasks (Chen et al., 2023; Fan et al., 2024; Liu et al., 2024; Glazer et al., 2024). While mathematics underpins logical reasoning, we explore applications requiring multi-step reasoning and domain-specific knowledge. Physics, a core natural science, provides a mathematical framework for modeling (Yung-kuo & Class, 1994; Jackson, 1998; Sakurai & Napolitano, 2020; Pathria & Beale, 2011), spanning both deterministic classical mechanics and probabilistic quantum physics. Its complexity and multi-step reasoning demands make it an ideal domain for evaluating models on advanced problem solving.

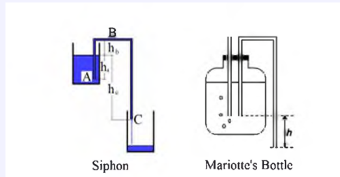
We present PHYSICS, a comprehensive and challenging benchmark designed to assess foundation models' physics problem-solving abilities. Unlike existing datasets that primarily consist of multiple-choice questions or focus on primary- to high school-level problems (Johannes Welbl, 2017; Lu et al., 2022; Yue et al., 2024b), PHYSICS is constructed from high-level, open-ended physics problems, specifically drawn from Physics PhD-qualifying exams[1]. In contrast to multiple-choice questions that may potentially allow models to exploit shortcuts or rely on answer recognition, PHYSICS is composed of comprehensive questions which minimizes the likelihood of bypassing the reasoning process and demands mastery of understanding theoretical concepts, complex inputs and the ability to integrate ideas from professional knowledge.

PHYSICS spans six core physics subjects: classical mechanics, quantum mechanics, thermodynamics and statistical mechanics, electromagnetism, atomic physics, and optics. These fields are selected as they encompass a diverse range of problem-solving techniques, requiring deep mathematical

*Equal contributions. Correspondence: {yilun.zhao, john.sous, arman.cohan}@yale.edu. Full author contributions in Appendix A.

[1]Physics PhD-qualifying exam constitutes problems covering the core areas of physics that every PhD student must master to qualify as a physicist. Traditionally, students need to pass this challenging exam before advancing to the research phase of their PhD program. These exams are typically difficult, with passing thresholds ranging from 40% to 70%.

Problem:
Below (left) is a diagram of a basic siphon from Wikipedia. Assume that the cross-section of the tube is much smaller than the dimensions of the water reservoir that the siphon is draining.
What is the speed of water flow out in terms of h_c? What is the maximum speed of smooth water flow out in terms of h_b? What is the maximum height h_b?



Solution:
Since the siphon cross-section is much smaller than the reservoir dimensions, we can treat the reservoir as essentially infinite: the speed at which the water level in the reservoir is decreasing is negligible compared to the flow speed through the tube. Set the origin of vertical height at the water level in the reservoir. Then, the constant in Bernoulli's equation is $\frac{1}{2}\rho(0)^2 + \rho g(0) + P_0 = P_0$.

Bernoulli states that this expression is the same at all points along the water flow. At C, this reads $\frac{1}{2}\rho v_c^2 - \rho g h_c + P_0 = P_0$, or $\boxed{v_c = \sqrt{2gh_c}}$. At B, this reads $\frac{1}{2}\rho v_b^2 + \rho g h_b + P_b = P_0$. We can drive P_b down to 0, but no further, or else cavitation occurs (bubble formation) and the flow will no longer be smooth. Setting $P_b = 0$ gives $\boxed{v_{\text{max}} = \sqrt{2\left(\frac{P_0}{\rho} - gh_b\right)}}$.

Requiring the above speed to be ≥ 0 yields the maximum height $\boxed{h_{b,\text{max}} = \frac{P_0}{\rho g}}$.

Figure 1: An example of a classical mechanics problem in PHYSICS. PHYSICS is a comprehensive benchmark for PhD-qualifying exam physics problem-solving with 1,297 expert-annotated problems.

modeling, multi-step logical reasoning, and theoretical integration. These settings makes PHYSICS ideal for evaluating the reasoning and analytical capabilities of foundation models.

To ensure fair and reliable evaluation, we develop a reliable automated evaluation system that is able to automate answer extraction, standardize mathematical expressions, and assess accuracy. Correctness is verified using `SymPy`, an open-source library for symbolic mathematics, for rule-based equivalence checking. When this fails, a GPT-4o-based assessment is applied.

We conduct an extensive evaluation across 33 frontier proprietary and open-source models. Our evaluation results show that even the best-performing model, o3-mini, achieves 59.9% accuracy. This reveals critical challenges facing these models, including their struggles with lengthy reasoning chains, reliance on incorrect assumptions, systematic errors, misunderstanding of images, and misinterpretation of problem statements. These failure modes persist across models, suggesting fundamental limitations in current models' capabilities. The performance gap between proprietary and open-source models further illuminates areas for improvements - most open-source models trail significantly, with even the most capable and tested model (Qwen2.5-Math-72B) reaching 32.2% accuracy. Extended chain-of-thought and self-reflective reasoning is also not sufficient to significantly boost performance, with DeepSeek-R1 achieving 44.3% accuracy. Through our detailed analyses and examining of specific failure patterns, we provide insights for current limitations of frontier models and for improving model developments for advanced reasoning in contextual, specialized domains.

We summarize our contributions as follows:

- We introduce a challenging benchmark featuring expert-annotated physics problems spanning six subfields. Our benchmark demands deep multi-step reasoning and theoretical knowledge integration, challenging frontier foundation models.
- We develop a robust automated evaluation framework that ensures precise and standardized assessment by leveraging `SymPy` and GPT-based evaluation, enhancing the reliability of model performance measurement.
- We conduct a comprehensive evaluation of both open-source and proprietary foundation models, systematically analyzing their strengths, weaknesses, and limitations in solving our benchmark.

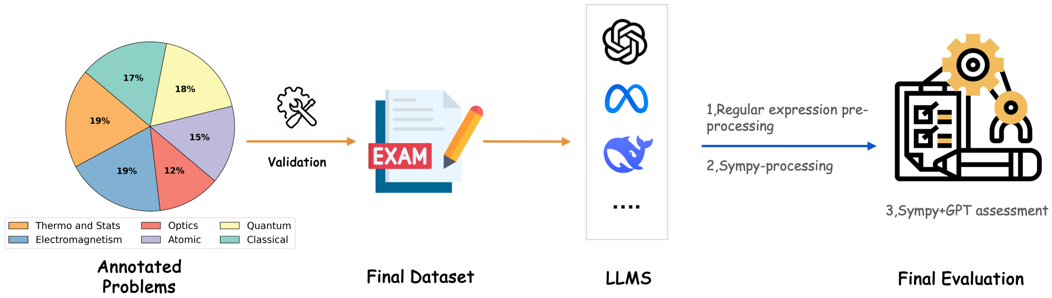


Figure 2: For the overall process, we begin by collecting annotated problems from annotators (§3.2), followed by validation to create a processed dataset. This dataset is then used to prompt models (§4.1). The responses from models undergo regular expression pre-processing and `SymPy`-based processing before final evaluation using an automated system (§4.2).

- We provide an in-depth analysis of different prompting techniques, Long CoT, failure case studies, and Retrieval-Augmented Generation (RAG)-based knowledge augmentation, offering valuable insights into guiding future improvements.

2 Related Work

As AI continues to advance in reasoning tasks, recent efforts such as Humanity's Last Exam (Phan et al., 2025) highlight the growing need for high-difficulty, domain-specific benchmarks. OpenAI's o1 models (OpenAI et al., 2024b) demonstrate significantly improved performance on existing math benchmarks, reinforcing the necessity for more challenging datasets that push foundation models toward deeper scientific reasoning. In response to this need, the evaluation of foundation models in mathematical reasoning has evolved through increasingly structured benchmarks. Early datasets like GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) assessed elementary to undergraduate-level problem solving, but state-of-the-art models now achieve near-perfect performance on these. More advanced benchmarks, such as HARDMATH (Fan et al., 2024), OlympiadBench (He et al., 2024), PutnamBench (Tsoukalas et al., 2024), and FrontierMath (Glazer et al., 2024), introduced university and olympiad-level problems, further pushing model capabilities.

This underscores the need to extend evaluation beyond mathematics into physics, where problem-solving involves real-world principles, physical laws, and multi-step derivations. The added complexity highlights the need for specialized benchmarks for advanced physics reasoning. As shown in Section 2, current scientific reasoning benchmarks, such as MMMU (Yue et al., 2024a), MMMU-Pro (Yue et al., 2024b), ScienceQA (Lu et al., 2022), and SciQ (Johannes Welbl, 2017), primarily focus on problems up to the high-school level. Meanwhile, benchmarks like SciEval (Sun et al., 2023) and SciKnowEval (Feng et al., 2024) include only a limited subset of physics, with a stronger focus on multiple choice questions. Thus, existing benchmarks lack the depth needed to evaluate AI's ability to solve advanced physics problems.

3 Physics Benchmark

We present Physics, a comprehensive benchmark for PhD-qualifying exam physics problems. Physics has the following key features: (1) **Comprehensive Subject Coverage**: The dataset spans six core fields of modern physics, encompassing a diverse range of fundamental and advanced topics, ensuring broad and extensive coverage of physics knowledge. (2) **Deep Multi-Step Reasoning**: Problems emphasize multi-step logical reasoning, requiring expertise in theoretical analysis, mathematical modeling, and complex problem-solving to assess advanced physics proficiency. (3) **Automated Verification for Rigorous Evaluation**: All solutions are verifiable using regular expressions and `SymPy`, enabling reliable correctness assessment and structured final answers for automated model benchmarking. (4) **Strict Data Quality Control**: Strict data quality control is enforced through a structured annotation process carried out by expert annotators, with all annotations meticulously cross-verified for accuracy and consistency.

Benchmark	Multi-modal	Size	Level	Question Type	Evaluation	Reasoning Steps
JEEBENCH(Arora et al., 2023)	✗	515	CEE	OE, MC	Rule-Based	-
MATH(Hendrycks et al., 2021)	✗	12,500	K12-Comp	OE	Rule-Based	-
HARDMATH(Fan et al., 2024)	✗	1,466	Graduate	OE	Rule + Model	-
GSM8K(Cobbe et al., 2021)	✗	8,500	K8	OE	Rule-Based	5.0
GPQA(Rein et al., 2024)	✗	227	Graduate	OE	Rule-Based	3.6
SciQ(Johannes Welbl, 2017)	✗	13,679	K4-K8	MC, OE	Rule-Based	-
SciEval(Sun et al., 2023)	✗	1657	-	OE, MC	Rule-Based	-
SCIBENCH(Wang et al., 2024)	✓	295	College	OE	Rule-Based	2.8
*MMMU*Yue et al. (2024a)	✓	443	College	OE, MC	Rule-Based	-
*MMMU-Pro*Yue et al. (2024b)	✓	3,460	College	MC	Rule-Based	-
SCIENCEQA(Lu et al., 2022)	✓	617	K1-K12	MC	Rule-Based	2.4
OlympiadBench(He et al., 2024)	✓	2334	Comp	OE	Rule-Based	3.7
PUTNAMBENCH(Tsoukalas et al., 2024)	✗	1692	College	OE	Rule-Based	-
PHYSICS	✓	1297	PhD-Qualifying	OE	Rule + Model	5.7

Table 1: Comparison of PHYSICS with other benchmarks. For **Level**, comp: Competition, COL: College, CEE: College Entrance Examination, K1-K12: Elementary and High School Level. For **Question Type**, OE: Open-ended Questions, MC: Multiple-choice Questions. **Reasoning Steps** are based on the statistics provided in the corresponding reference.

3.1 PROBLEM SELECTION

The benchmark dataset consists of 1,297 PhD-qualifying exam physics problems from publicly available sources, covering six key physics fields: classical mechanics, quantum mechanics, thermodynamics and statistical mechanics, electromagnetism, atomic physics, and optics. An example of a selected question is shown in Figure 1. These fields were selected because they encompass the fundamental principles essential for modeling and analyzing physical systems across various scales. Together, they offer a comprehensive evaluation of a model's capability to handle deterministic and probabilistic reasoning, maintain mathematical precision, and solve multi-step problems effectively.

3.2 DATA ANNOTATION

Our annotation process follows a structured pipeline to ensure high-quality and consistent data. A dedicated team of seven expert annotators, all proficient in physics, carried out the annotation process (detailed biographies are presented in Table 5). Before contributing, each annotator was required to pass a qualification test assessing their ability to accurately annotate complex physics problems. The problems were sourced from major textbooks and reputable online resources, with annotators ensuring that the selected questions adhered to copyright and licensing regulations.

For the problem annotation, annotators formatted problem statements using LaTeX , ensuring precision in mathematical notation, consistency in variable usage, and adherence to standard physics conventions. Key metadata, including problem conditions and underlying assumptions, were documented to enhance clarity and comprehension. For the solution annotation, each problem's solution was carefully structured with a step-by-step reasoning process, ensuring logical coherence and mathematical accuracy. Annotators incorporated theoretical justifications, derivations, and final answers in a structured format suitable for automated verification. Figures and diagrams were annotated to accurately represent physical concepts, insuring the clarity of problem-solving steps.

To assess problem difficulty, annotators were required to provide ratings based on the following:

- **Creativity**: The time required to identify the key steps necessary to solve the problem.
- **Complexity**: The number of attempts needed to arrive at a correct solution.

3.3 DATA QUALITY CHECK

To maintain mathematical rigor, annotators were explicitly required to focus on the accurate expression of equations, numerical constants, and symbolic representations. Special attention was given to ensuring that all mathematical formulations adhered to standard conventions and were free of typographical or conceptual errors.

To ensure the highest accuracy of annotations, a multi-step review mechanism was implemented. Each annotation underwent an independent verification process by the authors to confirm correctness and adherence to the intended guidelines. Ambiguities were solved through effective discussion between authors and annotators. Each problem underwent a secondary validation process, where a second annotator carefully reviewed the annotations to ensure accuracy and consistency. Any identified errors were corrected, and necessary refinements were made to align with the guidelines.

Category	Value
Dataset Overview	
Total Questions	1,297
Questions with Figures	298
Validation : Test Split	297 : 1,000
Hard : Regular Questions	523 : 774
Subject Distribution	
Number of Subjects	6
Atomic Physics	200
Electromagnetism	242
Classical Mechanics	221
Optics	158
Quantum Mechanics	236
Statistical Physics	240
Question Complexity	
Avg. Question Length (words)	83.7
Solution Statistics	
Avg. Solution Length (words)	234.75
Avg. Reasoning Steps	5.38

Table 2: Dataset statistics of PHYSICS.

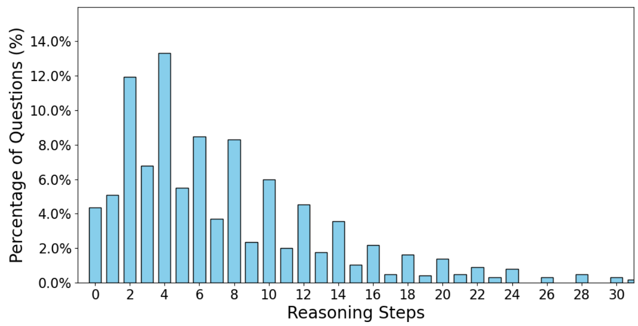


Table 3: Reasoning steps distribution.

3.4 DATA STATISTICS

Table 2 and Table 3 presents the key statistics of our dataset. It comprises a total of 1,297 questions, with 298 being multimodal format. To provide clear subsets for future researchers to train models and for testing, a total of 297 questions are randomly selected as the validation set for model development and validation, while the test set, which contains the remaining 1,000 questions, is reserved for standard evaluation to prevent data contamination. According to the difficulty assessment done by annotators, we select 523 questions as the HARD subset. This subset is designed to provide a more challenging benchmark for evaluating the capabilities of current frontier foundation models. Over the PHYSICS dataset, the average question length is 83.7 words, while solutions exhibit a significantly higher average length of 5.4 reasoning steps, with 24% of them containing more than 10 reasoning steps, reflecting the deep reasoning required for problem solving. These statistics highlight the dataset's potential to assess models on multi-step reasoning and precise answer generation in physics.

4 EXPERIMENT

This section discusses our experiment setup, main findings, and analysis.

4.1 EXPERIMENT SETUP

To systematically evaluate the capabilities of state-of-the-art models and establish clear reference points for future research, we conduct a comprehensive assessment across a diverse set of models including five proprietary models: GPT-4o (OpenAI et al., 2024a), Gemini-1.5-Pro (Team et al., 2024), Claude-3.5-Sonnet (Anthropic, 2024), o3-mini (OpenAI, 2025) and o1-mini (OpenAI et al., 2024b). Also 28 open-source models are evaluated, including DeepSeek-R1 (DeepSeek-AI et al., 2025), Llama 3.3 and Llama 3.1 series (Grattafiori et al., 2024), Qwen2, Qwen 2.5 and Qwen2.5-Math series (Qwen et al., 2025), QwQ (Team, 2024), Gemma2 series (Gemma Team, 2024), Command-R series (Cohere, 2024), Phi-4 (Abdin et al., 2024), Mistral, Mathstral and Mixtral series (Jiang et al., 2024), ChatGLM (GLM et al., 2024), Internlm3 (Cai et al., 2024), Yi-1.5 (AI, 2025), Aria (Li et al., 2025), InternVL2.5 (Chen et al., 2025) and Pixtral (Agrawal et al., 2024). Open-source models are evaluated using the vLLM pipeline (Kwon et al., 2023), while proprietary models are accessed via their official APIs. For text-only models, we tested on the subset of text-only questions. We adopt the Chain-of-Thought prompt used in MATH (Hendrycks et al., 2021) for model evaluation. The models are instructed to output the final answer in boxed format.

> Chain-of-Thought Prompt for LLM
>
> **Question:** {question}
>
> **Figure (For multimodal-format problem)**: {base64 processed image}
>
> **Prompt**: Answer the given question step by step. Begin by explaining your reasoning process clearly. Conclude by stating the final answer using the following format: "Provide the final answer at the end in LaTeX boxed format $\boxed{\text{final\_answer}}$." Think step by step before answering.

4.2 AUTOMATED EVALUATION SYSTEM

Model Response Processing. The model-generated response undergoes systematic processing to extract the final answer in a structured manner. First, we leverage Python's built-in re module to use regular expressions for searching and matching specific patterns within the model-generated response, thereby extracting the final boxed answer (`\boxed{}`). Following this, mathematical expressions are pre-processed to maintain consistency and validity. This includes standardizing notations, such as ensuring that fractions are formatted uniformly (*e.g.,* converting expressions like `\frac1b` to the explicit form `\frac{1}{b}`), thereby reducing variability in formatting. Furthermore, to ensure logical accuracy, only the core mathematical content relevant to the final answer is extracted for comparison. For example, in statements involving logical implications (*e.g.,* $\implies$), only the portion following the implication arrow is considered, as it represents the model's asserted conclusion. This extraction approach prevents extraneous elements from affecting correctness judgments.

Accuracy Evaluation. To rigorously assess the correctness of model responses, the framework utilizes a robust mathematical verification system. The processed LaTeX answers are parsed into symbolic representations using `sympy.parse_latex`, which enables direct mathematical equivalence checking through the function `sympy.is_equiv`. This ensures that two mathematically equivalent expressions are recognized as correct even if they appear in different algebraic forms. In scenarios where questions consist of multiple sub-answers, accuracy is computed as the fraction of correct responses relative to the total number of expected answers. For instance, if a question consists of five distinct parts and the model correctly answers three, the computed accuracy score would be 0.6, providing a fine-grained measure of performance. Specifically, GPT-4o is used to evaluate answers that rely on natural language explanations, ensuring correctness, relevance, and logical consistency. This approach is particularly important in physics, where conceptual understanding and domain-specific knowledge extend beyond mere mathematical calculations. Additionally, GPT-4o evaluation is employed as a fallback when symbolic computation (e.g., SymPy) returns False or fails to verify an answer. By integrating this backup method, we enhance the robustness and reliability of our accuracy evaluation system. Detailed prompts are shown in Appendix C.1

4.3 MAIN FINDINGS

Table 4 shows the model performance on the PHYSICS dataset. Our key findings are as follows:

PHYSICS challenges current models. Our benchmark presents significant challenges for current foundation models. Even the most advanced proprietary models like o3-mini achieve only 59.9% of accuracy, while other mainstream models such as GPT-4o and Gemini-1.5-Pro achieve only approximately 37% accuracy, highlighting a substantial gap in their ability to handle complex mathematical and conceptual reasoning.

In contrast, human performance on similar tasks is significantly higher. We randomly select 10 tasks for each human expert. Given adequate time, human experts typically achieve scores ranging from 70% to 80%. This disparity highlights the limitations of current models in reasoning through multi-step physics problems.

Proprietary Model Performance Analysis. Proprietary models, such as GPT-4o and Gemini-1.5-Pro, achieve an accuracy of approximately 37% on our benchmark. While these state-of-the-art

	Test Set Performance						Overall	
Model	**AMO**	**E&M**	**CM**	**Opt.**	**QM**	**Stats.**	**Val**	**Test**
Proprietary Models								
o3-mini	52.4	64.9	59.8	51.5	66.0	60.0	55.0	59.9
o1-mini	45.4	41.8	41.9	40.6	44.3	48.0	44.1	43.6
Gemini-1.5-pro†	35.5	40.2	31.5	32.2	44.5	43.7	35.3	38.4
GPT-4o†	35.3	44.1	33.4	23.4	33.8	45.0	34.7	36.7
Claude-3.5-Sonnet†	37.2	34.8	27.6	35.5	35.1	38.4	31.7	34.7
Open-Source Models								
DeepSeek-R1	37.0	48.6	38.3	43.1	44.5	51.5	44.2	44.3
Qwen2.5-Math-72B	27.0	34.8	27.3	27.4	36.2	37.0	38.5	32.2
Llama-3.3-70B	28.2	35.8	27.9	17.2	31.4	41.3	34.3	31.5
phi-4	32.8	33.0	19.8	27.2	23.4	35.2	28.7	29.1
Qwen2.5-72B	28.8	30.9	23.0	25.4	27.4	33.2	31.5	28.7
Qwen2.5-32B	25.5	27.5	19.4	20.8	24.7	41.1	23.3	27.6
Mistral-Small-24B	19.1	29.5	19.6	17.6	15.2	28.4	25.1	21.8
Qwen2.5-7B	21.8	28.1	11.2	18.7	17.4	22.1	20.9	20.4
Qwen2.5-14B	23.8	19.7	14.1	12.3	13.5	28.2	25.3	19.6
Gemma-2-27b	14.3	19.0	16.2	13.4	18.4	25.9	21.7	18.3
Yi-1.5-34B	11.0	15.4	18.0	13.2	19.6	25.2	25.3	17.4
Qwen2.5-Math-1.5B	13.3	14.8	16.5	16.2	17.2	19.5	15.1	16.4
InternVL2-5-38B†	15.3	12.5	12.5	7.7	18.0	23.1	16.7	15.3
Aria†	13.0	14.0	14.2	11.7	9.7	14.4	12.7	12.9
QwQ-32B-Preview	16.7	7.5	10.1	11.2	10.6	14.8	12.4	12.1
Gemma-2-9b	9.4	8.2	9.1	16.5	12.1	16.9	15.2	11.9
Mistral-7B	10.1	10.4	5.1	13.7	11.6	17.6	12.6	11.7
Llama-3.1-8B	8.4	17.4	6.8	14.7	7.4	16.1	9.1	11.7
Mathstral-7B	7.3	10.0	12.0	9.6	8.2	17.6	12.0	10.8
c4ai-command-r-v01	2.0	7.8	7.5	3.8	7.5	11.4	6.8	7.0
DeepSeek-R1-Distill-Qwen-32B	9.1	5.4	4.8	9.8	2.3	10.2	7.1	6.8
Gemma-2-2b	6.6	6.2	3.9	10.3	3.9	7.3	6.1	6.1
Qwen2-VL-72B†	11.8	3.5	4.6	4.0	2.9	4.2	4.5	5.0
Internlm3-8b	1.8	4.6	4.7	3.2	4.0	9.2	4.1	4.8
DeepSeek-vl2-small†	3.1	1.8	1.8	4.5	0.0	0.3	4.8	1.7
THUDM-chatglm3-6b	0.9	2.3	0.0	0.7	0.9	2.0	0.9	1.2
Qwen2.5-Math-7B	1.4	1.7	0.0	2.1	0.0	1.5	1.9	1.0
DeepSeek-math-7b-rl	0.7	0.0	0.0	1.5	0.0	0.6	0.9	0.4

Table 4: Performance comparison across tasks. †: These models are equipped with multi-model abilities. Problems with images are also tested on these models. **Abbreviations:** AMO (Atomic Physics) — E&M (Electromagnetism) — CM (Classical Mechanics) — Opt. (Optics) — QM (Quantum Mechanics) — Stats. (Theromodynamics and Statistical Physics). The models are ranked by average test set performance.

models exhibit strong reasoning capabilities and a solid grasp of problem comprehension, their performance underscores a significant gap in solving PhD-qualifying exam physics problems. Notably, frontier models such as o3-mini, which incorporate system-2 reasoning, demonstrate a substantial leap in performance compared to other proprietary models. This improvement underscores the effectiveness of Chain-of-Thought (CoT) reasoning and deep-thinking training, further emphasizing the importance of structured reasoning paradigms in enhancing AI capabilities on challenging tasks.

Open-Source Model Performance Analysis. For open-source models, we do observe DeepSeek-R1 achieved the highest score of 44.3% in accuracy, due to its advanced training in CoT and reasoning abilities. Other state-of-the-art open-source models, such as Llama-3.3-70B, Phi-4, and Qwen-2.5-72B, achieved an accuracy of approximately 30%. However, beyond these leading models, the majority of open-source models exhibit significantly lower performance, with accuracy lower than 20%. This performance gap highlights a substantial disparity between open-source models and their proprietary counterparts. While open-source models continue to make progress, their effectiveness in complex reasoning tasks remains limited compared to state-of-the-art proprietary models. This underscores the need for further advancements in the open-source AI community, particularly in enhancing multimodal and expert-level reasoning capabilities.

4.4 Qualitative Analysis

We conduct a detailed case study on the top-performing models. Specifically, we focus on GPT-4o, Llama3.3-70B, and DeepSeek-R1 to gain deeper insights into their strengths and weaknesses.

Analysis of Multimodal Models. PHYSICS includes problems that require image understanding. We conduct case studies on specific tasks where interpreting visual information is essential for problem solving. The experiments reveal errors such as failures in interpreting expressed spatial information or misunderstanding the relation between objects in the image. These issues often lead to incorrect information in the models' reasoning processes. An example is provided in Appendix E.

Inability to integrate Professional Knowledge. One issue with foundation models when reasoning about physics problems is their tendency to misinterpret or overlook fundamental real-world principles that are not explicitly stated in the problem. This limitation indicates their inability to integrate essential background knowledge, often leading to incorrect assumptions or flawed conclusions. Addressing this gap requires models to develop a deeper understanding of physical systems and incorporate common knowledge when solving such problems. An example is shown in Appendix D.2.

Non-Existent Assumptions. Another frequent issue is the introduction of extraneous conditions that were not explicitly provided in the original question. These additional assumptions alter the problem scope, leading to incorrect conclusions and reducing the reliability of the generated solution. This reflects the need for models with the ability to reason within the given constraints while ensuring that no unwarranted premises are introduced, thereby preserving the accuracy and relevance of the solution. An example is shown in Appendix D.3.

Calculation Errors in Complex Equations. Despite demonstrating strong reasoning abilities, models can still struggle with computational accuracy when dealing with intricate mathematical expressions. Errors frequently emerge in multi-step calculations, particularly those involving symbolic manipulations and algebraic simplifications. These mistakes undermine the reliability of model-generated responses, emphasizing the need for improved numerical precision and robust verification mechanisms. An example is shown in Appendix D.4.

Misunderstanding Question. An occasional issue in model evaluations is the failure to correctly interpret the problem statement, leading to irrelevant or inaccurate responses. This issue does not arise often. However, we find that smaller-scale models are more prone to misinterpretation. This manifests in several ways: misidentifying key variables, overlooking critical constraints, or applying incorrect problem-solving frameworks. Such misunderstandings can result in fundamentally flawed reasoning chains that diverge from the intended solution path. Additionally, models may misinterpret ambiguous phrasing, leading to incorrect assumptions that were not implied by the question. An example is shown in Appendix D.5.

We believe that addressing the aforementioned shortcomings will be crucial in developing foundation models capable of handling complex reasoning tasks in physics and other scientific domains.

4.5 Analysis of Different Prompting Methods

To evaluate the impact of self-reflection on model performance, we compare Chain-of-Thought and Self-Reflection (Renze & Guven, 2024) prompting methods (prompts provided in Appendix C.2) with the four best-performing models on our benchmark. Models prompted with self-reflection generally demonstrate improved reasoning consistency and accuracy compared to CoT prompting (Figure 3). These findings highlight the broad effectiveness of self-reflection across different model scales, as it enhances problem-solving accuracy, making foundation models more reliable and efficient across diverse tasks and architectures.

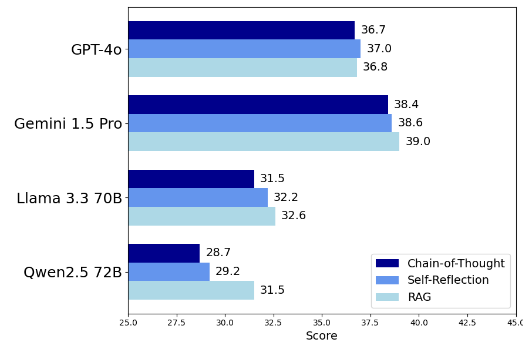


Figure 3: Comparison between different methods.

4.6 RAG-based Knowledge Augmentation

We next propose the use of Retrieval-Augmented Generation (RAG) to address performance limitations of foundation models in integrating professional knowledge for physics. To determine if access to external knowledge can mitigate this weakness, we compare the accuracy and reliability of four top-performing models on our benchmark both with and without RAG augmentation. Our RAG implementation utilizes SerpAPI's Google Search, with models prompted (see Appendix C.3) to autonomously formulate search queries based on the presented physics questions. The context of top-5 search results is concatenated to the original query and fed into the models. As shown in Figure 3, the RAG setting improves performance across all tested models. These findings underscore the potential of retrieval-based augmentation in enhancing model reasoning capabilities. Highlighting the importance of external knowledge integration in addressing gaps in foundational models' domain-specific understanding.

4.7 Analysis of Models with Long CoT Ability

Models trained for stronger reasoning abilities, such as o3-mini, o1-mini, and DeepSeek-R1, exhibit significant improvements in physics reasoning. Their enhanced capability for long Chain-of-Thought (CoT) reasoning enables these foundation models to outperform others, demonstrating the effectiveness of prioritizing advanced AI reasoning abilities in training for solving physics problems. A common issue observed in DeepSeek-R1 and QwQ-32B which encourage self-reflection (we did not analyze OpenAI's o1-series models as the full reasoning outputs are not available) is the tendency to generate overly extensive reasoning chains. While self-reflection thinking can be beneficial, it often results in an unnecessarily prolonged thinking process, frequently exceeding the 10240-token limit. This leads to incomplete answers or excessive computational overhead, diminishing the overall effectiveness of the response. Examples are shown in Appendix D.1.

5 Conclusion

We introduce PHYSICS, a benchmark designed to evaluate foundation models on PhD-qualifying exam physics problems, comprising 1,297 questions across six core subfields. Our results show that even leading proprietary models, o3-mini, achieve only 59.9% accuracy, underscoring fundamental weaknesses in scientific reasoning, conceptual understanding, and mathematical precision. Through a detailed failure analysis, we identify five key failure patterns: (1) Inability to integrate professional knowledge, (2) reliance on incorrect assumptions, (3) difficulties in handling multimodal data, (4) calculation errors in multi-step reasoning, and (5) misunderstanding questions. Various prompting methods and RAG-based knowledge augmentation are also explored. These methods demonstrate potential for improvement but do not fully bridge the gap in expert-level physics problem-solving. These findings emphasize the need for enhanced reasoning frameworks, improved mathematical precision, and effective integration of physics knowledge sources to bridge the gap. PHYSICS establishes a comprehensive evaluation framework, providing a foundation for measuring progress in scientific reasoning and provides insights for future AI model development in specialized domains.

ACKNOWLEDGMENTS

This work is supported in part by the NVIDIA Academic Grant Program. We are grateful to Google TRC program for providing computing resources and Together AI for granting LLM API credits.

LIMITATIONS

While PHYSICS serves as a valuable benchmark for AI-driven physics problem solving, it has some limitations. One limitation is its reliance on automated evaluation using SymPy for rule-based checking and GPT-based assessments. While effective, this approach may fail to recognize equivalent solutions and introduce subjectivity. Incorporating expert human review could enhance evaluation accuracy. Moreover, although PHYSICS is extensive, comprising 1,297 questions, it does not fully encompass the breadth of physics. More advanced and interdisciplinary topics remain underrepresented. Expanding the dataset to include a wider range of complex topics would enhance its comprehensiveness. We encourage future research to address these gaps by incorporating more advanced and diverse areas of physics.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, and et al. Pixtral 12b, 2024. URL https://arxiv.org/abs/2410.07073. ICML Submission.

01. AI. Yi: Open foundation models by 01.ai, 2025. URL https://arxiv.org/abs/2403.04652.

Anthropic. Introducing claude 3.5 sonnet, June 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet. Accessed: 2024-06-21.

Daman Arora, Himanshu Gaurav Singh, and Mausam . Have LLMs advanced enough? a challenging problem solving benchmark for large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=YHWXlESeS8.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, and et al. Internlm2 technical report, 2024.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. TheoremQA: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7889–7901, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.489. URL https://aclanthology.org/2023.emnlp-main.489/.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, and et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. URL https://arxiv.org/abs/2412.05271. ICML Submission.

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. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Cohere. Command r and command r+ model card, 2024. URL https://docs.cohere.com/docs/responsible-use#usage-notes. Accessed: 2025-01-29.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, and et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948. ICML Submission.

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*, 2024. URL https://openreview.net/forum?id=gt6prlTEGL.

Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. Sciknoweval: Evaluating multi-level scientific knowledge of large language models, 2024.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie S tewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai, 2024. URL https://arxiv.org/abs/2411.04872.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, and et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024. URL https://arxiv.org/abs/2406.12793. ICML Submission.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783. ICML Submission.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

John David Jackson. *Classical Electrodynamics*. Wiley, 3rd edition, 1998.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, and et al. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088. ICML Submission.

Matt Gardner Johannes Welbl, Nelson F. Liu. Crowdsourcing multiple choice science questions. 2017.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, Chongyan Zhu, Xiaoyi Ren, Chao Li, Yifan Ye, Peng Liu, Lihuan Zhang, Hanshu Yan, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model, 2025. URL https://arxiv.org/abs/2410.05993.

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

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

OpenAI. Openai o3-mini system card, 2025. URL `https://openai.com/index/o3-mini-system-card/`. Accessed: February 13, 2025.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, and et al. Gpt-4o system card, 2024a. URL `https://arxiv.org/abs/2410.21276`.

OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, and et al. Openai o1 system card, 2024b. URL `https://arxiv.org/abs/2412.16720`.

R.K. Pathria and Paul D. Beale. *Statistical Mechanics*. Elsevier, 3rd edition, 2011.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, and et al. Humanity's last exam, 2025. URL `https://arxiv.org/abs/2501.14249`.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, and et al. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`. ICML Submission.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL `https://openreview.net/forum?id=Ti67584b98`.

Matthew Renze and Erhan Guven. Self-reflection in llm agents: Effects on problem-solving performance, 2024. URL `https://arxiv.org/abs/2405.06682`.

J. J. Sakurai and Jim Napolitano. *Modern Quantum Mechanics*. Cambridge University Press, 3rd edition, 2020.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. Scieval: A multi-level large language model evaluation benchmark for scientific research. *arXiv preprint arXiv:2308.13149*, 2023.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL `https://arxiv.org/abs/2403.05530`.

Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown, November 2024. URL `https://qwenlm.github.io/blog/qwq-32b-preview/`.

George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings, Amitayush Thakur, and Swarat Chaudhuri. Putnambench: A multilingual competition-mathematics benchmark for formal theorem-proving. In *AI for Math Workshop @ ICML 2024*, 2024. URL `https://openreview.net/forum?id=vqW1VRFeVP`.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=bq1JEgioLr`.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, June 2024a.

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

Lim Yung-kuo and The Physics Coaching Class. *Problems and Solutions on Thermodynamics and Statistical Mechanics*. World Scientific, 1994. Major American Universities Ph.D. Qualifying Questions and Solutions.

A AUTHOR CONTRIBUTIONS

- Project leads: Kaiyue Feng and Yilun Zhao
- Initial project idea: Kaiyue Feng, Yilun Zhao, John Sous, and Arman Cohan.
- Data annotation task management and validation: Kaiyue Feng, Yilun Zhao, and Tianyu Yang.
- Experiments: Kaiyue Feng, Yilun Zhao, and Yixin Liu.
- Writing: Kaiyue Feng and Yilun Zhao led the writing, with editing contributions from all authors.
- Mentorship and advising: Chen Zhao, John Sous, and Arman Cohan.

B ANNOTATION BIOGRAPHY

This section presents an overview of the biography of annotators, including their qualifications, majors, and assigned subjects.

ID	Year	Major	Assigned Subject(s)
1	2nd Year Undergraduate	Physics	Classical Mechanics, Electromagnetism
2	3rd Year Undergraduate	Physics	Quantum Mechanics, Optics
3	2nd Year Undergraduate	Theoretical Physics	Quantum Mechanics, Thermodynamics and Statistical Mechanics
4	3rd Year Undergraduate	Applied Physics	Thermodynamics and Statistical Mechanics, Atomic Physics
5	2nd Year Undergraduate	Engineering Physics	Electromagnetism, Classical Mechanics
6	3rd Year Undergraduate	Physics	Thermodynamics and Statistical Mechanics, Atomic Physics
7	2nd Year Undergraduate	Astrophysics	Classical Mechanics, Optics

Table 5: Biographies of 7 annotators involved in the Physics benchmark construction

C PROMPTS

This section documents the various prompts designed and utilized in our LLM experiments. Each prompt was crafted to elicit specific types of responses from the LLMs being evaluated.

C.1 MATH EQUIVALENCY PROMPT

Prompt for LLM

Question: {question}

Image: {base64 processed image}

Task: Compare the following LaTeX expressions and determine if the numerical content is equivalent.

Expression 1: {expr1}
Expression 2: {expr2}
Return **True** if they are mathematically equivalent; otherwise, return **False**. Focus only on numerical and mathematical content. If the expressions involve multiple-choice answers (e.g., A, B, C, D), compare only the letter choices.

C.2 SELF-REFLECTION PROMPT

Prompt for LLM

Initial Answer Step:

```
"You are an AI expert specializing in answering advanced
physics questions. Think step by step and provide
solution and final answer. Provide the final answer
at the end in LaTeX boxed format \[ \boxed{} \].
Example: \[ \boxed{ final_answer} \]"
```

Refinement Step:

```
# Put the initial answer into the context of the conversation
and ask the LLM to self-check and improve the answer
messages.append({
    "role": "assistant",
    "content": first_response
})
messages.append({
    "role": "user",
    "content": (
        "Please check your previous answer carefully.
        Identify any mistakes and refine your final answer.
        Provide the revised answer at the end
        in LaTeX boxed format \[ \boxed{} \]."
    )
})
```

C.3 RAG PROMPT

Prompt for LLM

Prompt:"You are a physics expert skilled in using search engines to find relevant information. Based on the given physics problem, think about the best way to query Google. Generate up to 3 relevant search queries that focus on key concepts, formulas, or theories. Do not search the entire question; instead, extract essential elements "

Task: "Generate relevant Google search queries for the following physics problem:{question_text} "

D Error Analysis

D.1 Excessively Long Reasoning Chains

Question:
It can be shown that the electric field inside a dielectric sphere which is placed inside a large parallel-plate capacitor is uniform (the magnitude and direction of E_0 are constant). If the sphere has radius R and relative dielectric constant $K_e = \varepsilon/\varepsilon_0$, find E at point p on the outer surface of the sphere (use polar coordinates R, θ). Determine the bound surface charge density at point p.

Solutions:
The electric field inside the sphere is a uniform field E_0. The field at point p of the outer surface of the sphere is

$$E = E_r e_r + E_t e_\theta,$$

using polar coordinates. Similarly, E_0 may be expressed as

$$E_0 = E_0 \cos\theta e_r - E_0 \sin\theta e_\theta.$$

From the boundary conditions for the electric vectors at p, we obtain

$$\varepsilon E_0 \cos\theta = \varepsilon_0 E_r, \quad -E_0 \sin\theta = E_t.$$

Hence,

$$E = K_e E_0 \cos\theta e_r - E_0 \sin\theta e_\theta.$$

The bound surface charge density at point p is $\sigma_b = \mathbf{P} \cdot e_r$, where $\mathbf{P}$ is the polarization vector. As

$$\mathbf{P} = (\varepsilon - \varepsilon_0) E_0,$$

we find

$$\sigma_p = (\varepsilon - \varepsilon_0) E_0 \cos\theta = \varepsilon_0 (K_e - 1) E_0 \cos\theta.$$

Deepseek-R1

Okay, so I have this problem about a dielectric sphere inside a parallel-plate capacitor. The question says that the electric field inside the sphere is uniform with magnitude E_0. I need to find the electric field E at a point p on the outer surface of the sphere using polar coordinates R and θ. Also, determine the bound surface charge density at point p.

Hmm, let me start by recalling some concepts from electromagnetism related to dielectrics. When you place a dielectric material into an external electric field, it gets polarized. This polarization creates bound charges: volume bound charges and surface bound charges. The key here might be understanding how these bound charges affect the overall electric field both inside and outside the dielectric.

Since the sphere is placed within a larger system—a parallel-plate capacitor—the initial setup probably involves applying an external electric field, but because the sphere's presence modifies the field around itself.

wait, actually, the problem states that the electric field inside the sphere is already given as uniform, E_0. So maybe we don't need to worry too much about calculating E_0; instead, our focus should be on finding the electric field just outside the sphere and then determining the surface charge density there.

I remember that for linear isotropic dielectrics, the relation between the electric displacement $\mathbf{D}$ and the electric field $\mathbf{E}$ is given by

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P},$$

where $\mathbf{P}$ is the polarization. And since the relative permittivity K_e is defined as $\varepsilon/\varepsilon_0$, perhaps we'll use that somewhere.

Also, Gauss's Law comes into play when dealing with symmetric systems like spheres. But wait, the original capacitor plates create a uniform field, right? However, placing a spherical object would perturb that field—so near the sphere, especially externally, the field won't be exactly uniform anymore.

But hold on... This extends to the end

Error Category: Excessively Long Reasoning Chains.

Error Reason: The model did show reasoning process and do seem to be mimicking human's evaluation process. But the reasoning process is way too long.

D.2 INABILITY TO INTEGRATE PROFESSIONAL KNOWLEDGE422

Question:
A particle of mass m moves one-dimensionally in the oscillator potential

$$V(x) = \frac{1}{2}m\omega^2 x^2.$$

In the nonrelativistic limit, where the kinetic energy T and momentum p are related by

$$T = \frac{p^2}{2m},$$

the ground state energy is well known to be $\frac{1}{2}\hbar\omega$.
Allow for relativistic corrections in the relation between T and p and compute the ground state level shift ΔE to order $\frac{1}{c^2}$ (c = speed of light).

Solutions:
In relativistic motion, the kinetic energy T is given by

$$T \equiv E - mc^2 = \sqrt{m^2c^4 + p^2c^2} - mc^2.$$

Expanding using a binomial approximation:

$$\begin{aligned} T &= mc^2\left(1 + \frac{p^2}{m^2c^2}\right)^{\frac{1}{2}} - mc^2 \\ &\approx mc^2\left(1 + \frac{p^2}{2m^2c^2} - \frac{p^4}{8m^4c^4}\right) - mc^2 \\ &= \frac{p^2}{2m} - \frac{p^4}{8m^3c^2}. \end{aligned}$$

To order $\frac{1}{c^2}$, the term $-\frac{p^4}{8m^3c^2}$ is treated as a perturbation. Then, the energy shift of the ground state is

$$\Delta E = \left\langle -\frac{p^4}{8m^3c^2} \right\rangle = \int_{-\infty}^{\infty} \phi_0^* \left(-\frac{p^4}{8m^3c^2}\right)\phi_0 \, dx.$$

Substituting the harmonic oscillator ground state wavefunction:

$$\psi_0(x) = \left(\frac{m\omega}{\pi\hbar}\right)^{1/4} \exp\left[-\frac{m\omega}{2\hbar}x^2\right],$$

and computing the expectation value,

$$\Delta E = -\frac{15}{32}\frac{(\hbar\omega)^2}{mc^2}.$$

GPT-4o

To compute the relativistic correction to the ground state energy of the harmonic oscillator, we need to account for the relativistic relationship between the kinetic energy T and the momentum p. The relativistic kinetic energy is given by:

$$T = \sqrt{p^2c^2 + m^2c^4} - mc^2.$$

In the nonrelativistic limit, this can be expanded in powers of $\frac{p^2}{m^2c^2}$ using a binomial expansion:

$$T = \frac{p^2}{2m} - \frac{p^4}{8m^3c^2} + \mathcal{O}\left(\frac{1}{c^4}\right).$$

The first term, $\frac{p^2}{2m}$, is the usual nonrelativistic kinetic energy, while the second term, $-\frac{p^4}{8m^3c^2}$, represents the leading relativistic correction. This correction modifies the Hamiltonian of the system.
Step 1: Relativistic Hamiltonian (correct part ommited)
Step 2: Ground State of the Harmonic Oscillator (correct part ommited)
Step 3: Expectation Value of p^4 To compute the first-order correction to the energy, we need the expectation value of p^4 in the ground state. Using $\langle p^4\rangle = (\langle p^2\rangle)^2$, we have:

$$\langle p^4\rangle = (\hbar m\omega)^2.$$

The GPT solution assumes

$$\langle p^4\rangle = (\langle p^2\rangle)^2,$$

which is incorrect for a quantum harmonic oscillator. Instead, the correct value is:

$$\langle p^4\rangle = \frac{3}{4}(\hbar m\omega)^2.$$

Step 4: First-Order Energy Correction (rest ommited)
Final Answer The relativistic correction to the ground state energy is:

$$\Delta E = -\frac{\hbar^2\omega^2}{8mc^2}.$$

Thus, the corrected ground state energy is:

$$E_0 = \frac{1}{2}\hbar\omega - \frac{\hbar^2\omega^2}{8mc^2}.$$

$$\boxed{\Delta E = -\frac{\hbar^2\omega^2}{8mc^2}}$$

Error Category: Error Assumptions

Error Reason: The model demonstrated a complete reasoning process but made an error assumption when calculating, leading to a incorrect answer

D.3 Non-Existent Assumptions

Questions:
An astronaut places a turned-on flashlight in space. The flashlight is spin-stabilized, meaning it rotates about its own axis. The flashlight emits photons, which exert radiation pressure and act as a small thrust (photon rocket effect). The batteries last for two hours. We need to determine the additional velocity gained by the flashlight due to photon emission.

Solutions:
Suppose the flashlight bulb is located at the focus of a paraboloid reflector so that almost the entire light output goes out in one direction. If the power of the flashlight is N watts and the time interval it is turned on is t, the total energy of the photons emitted is $E = Nt$. If the orientation of the flashlight does not change, it will gain a momentum

$$mv = \frac{E}{c} = \frac{Nt}{c},$$

or an additional speed

$$v = \frac{Nt}{mc},$$

m being the mass of the flashlight, since a photon of energy ϵ has a momentum $\frac{\epsilon}{c}$.
For example, if $N = 1$ W, $m = 0.3$ kg, $t = 2$ hours,

$$v = \frac{1 \times 2 \times 3600}{0.3 \times 3 \times 10^8} = 8 \times 10^{-5} \text{ m/s}.$$

GPT-4o

Incorrect Solution

The force exerted by photon emission is given by:

$$F = \frac{P}{c} \tag{1}$$

where:

- P is the power of the flashlight,
- $c = 3 \times 10^8$ m/s is the speed of light.

Assuming a typical flashlight power of $P = 10$ W, we calculate:

$$F = \frac{10}{3 \times 10^8} = 3.33 \times 10^{-8} \text{ N} \tag{2}$$

Using Newton's second law:

$$a = \frac{F}{m} \tag{3}$$

Assuming the flashlight has a mass of $m = 1$ kg:

$$a = \frac{3.33 \times 10^{-8}}{1} = 3.33 \times 10^{-8} \text{ m/s}^2 \tag{4}$$

The time for which the flashlight emits light is:

$$t = 2 \times 3600 = 7200 \text{ s} \tag{5}$$

Thus, the velocity change is:

$$\Delta v = at = (3.33 \times 10^{-8}) \times (7200) \tag{6}$$

$$\Delta v = 2.4 \times 10^{-4} \text{ m/s} = 0.24 \text{ mm/s} \tag{7}$$

Mistake in the Solution

This answer incorrectly assumes that the rotation of the flashlight does not affect its propulsion. However, if the flashlight is spinning, the direction of photon emission changes continuously, meaning the net thrust could average out to zero. The correct approach must account for the rotational dynamics before concluding the velocity gain.

D.4 CALCULATION ERRORS IN COMPLEX EQUATIONS

Questions:
A nonrelativistic particle of mass m undergoes one-dimensional motion in the potential

$$V(z) = -g[\delta(x-a) + \delta(x+a)]$$

where g > 0 is a constant and $\delta(x)$ is the Dirac delta function. Find the ground-state energy eigenfunction and obtain an equation which relates the corresponding energy eigenvalue to the constant g.

Solutions:
Since $V(x) = V(-x)$, the energy eigenfunctions have definite parity. The ground state has even parity, $\psi(-x) = \psi(x)$. It is a bound state and its energy is negative, $E < 0$.
For $x \geq 0$, the Schrödinger equation is

$$\left[-(\hbar^2/2m)\, d^2/dx^2 - g\delta(x-a)\right] \psi(x) = E\psi(x),$$

whose solutions for $x \neq a$ are $\psi \sim \exp(fkx)$, where $k = \sqrt{2mE}/\hbar$.
With the condition that the wave function remains finite as $x \to \infty$ and has even parity, we obtain

$$\psi(x) = \begin{cases} Ae^{-kx}, & x > a, \\ B\cosh(kx), & 0 \leq x \leq a. \end{cases}$$

The continuity of ψ at $x = a$ requires that $A = Be^{ka}\cosh(ka)$. Thus

$$\psi(x) = \begin{cases} Be^{ka}\cosh(ka)e^{-kx}, & x > a, \\ B\cosh(kx), & 0 \leq x \leq a. \end{cases}$$

Normalization $\int_0^a \psi^2 dx + \int_a^\infty \psi^2 dx = \frac{1}{2}$ gives

$$B = \left(\frac{e^{2ka}}{2k} + \frac{1+2ka}{2k}\right)^{-1/2}$$

At $x = a$, there is a discontinuity of the first differential of the wave function (cf Problem 1024):

$$\psi'(a+) - \psi'(a-) = -(2mg/\hbar^2)\psi(a).$$

Substitution of ψ gives

$$k\left[1 + \tanh(ka)\right] = 2mg/\hbar^2,$$

which is to be satisfied by k. By symmetry the wave function for the entire space is

$$\psi(x) = \begin{cases} Be^{ka}\cosh(ka)e^{-k|x|}, & |x| > a, \\ B\cosh(kx), & |x| \leq a. \end{cases}$$

Llama-3.3-70B

Error in Calculation Example

From the Schrödinger equation, we use the boundary condition at $x = a$:

$$\psi'(a+) - \psi'(a-) = -\frac{2mg}{\hbar^2}\psi(a). \tag{8}$$

The derivative of the wave function in the two regions is:
For $x > a$:

$$\psi(x) = Be^{ka}\cosh(ka)e^{-kx}, \tag{9}$$

$$\psi'(x) = -kBe^{ka}\cosh(ka)e^{-kx}. \tag{10}$$

Evaluating at $x = a$:

$$\psi'(a+) = -kBe^{ka}\cosh(ka)e^{-ka} = -kB\cosh(ka). \tag{11}$$

For $0 \leq x \leq a$:

$$\psi(x) = B\cosh(kx), \tag{12}$$

$$\psi'(x) = kB\sinh(kx). \tag{13}$$

Evaluating at $x = a$:

$$\psi'(a-) = kB\sinh(ka). \tag{14}$$

Substituting into the boundary condition,

$$-kB\cosh(ka) - kB\sinh(ka) = -\frac{2mg}{\hbar^2}B\cosh(ka). \tag{15}$$

Cancelling B and rearranging:

$$-k(\cosh(ka) + \sinh(ka)) = -\frac{2mg}{\hbar^2}\cosh(ka). \tag{16}$$

Error: Incorrect Identity Used

$$\cosh(x) + \sinh(x) = 0 \quad \text{(Incorrect!)} \tag{17}$$

This incorrect step leads to:

$$-k(0) = -\frac{2mg}{\hbar^2}\cosh(ka), \tag{18}$$

Error: Invalid Conclusion This suggests either $g = 0$ (which contradicts the assumption $g > 0$) or $\cosh(ka) = 0$, which is impossible for real ka.

D.5 MISUNDERSTANDING QUESTION

Questions:
Consider the static magnetic field given in rectangular coordinates by

$$\mathbf{B} = B_0(x\hat{x} - y\hat{y})/a.$$

(a) Show that this field obeys Maxwell's equations in free space.
(b) Sketch the field lines and indicate where filamentary currents would be placed to approximate such a field.
(c) Calculate the magnetic flux per unit length in the $\hat{z}$-direction between the origin and the field line whose minimum distance from the origin is R.
(d) If an observer is moving with a non-relativistic velocity $\mathbf{v} = v\hat{z}$ at some location (x, y), what electric potential would he measure relative to the origin?
(e) If the magnetic field $B_0(t)$ is slowly varying in time, what electric field would a stationary observer at location (x, y) measure?

Solutions:
(a)

$$\begin{aligned}
\nabla \cdot \mathbf{B} &= \left(\hat{x}\frac{\partial}{\partial x} + \hat{y}\frac{\partial}{\partial y} + \hat{z}\frac{\partial}{\partial z}\right) \cdot \left[\frac{B_0}{a}(x\hat{x} - y\hat{y})\right] \\
&= \frac{B_0}{a}(\hat{x} \cdot \hat{x} - \hat{y} \cdot \hat{y}) = 0, \\
\nabla \times \mathbf{B} &= \left(\hat{x}\frac{\partial}{\partial x} + \hat{y}\frac{\partial}{\partial y} + \hat{z}\frac{\partial}{\partial z}\right) \times \left[\frac{B_0}{a}(x\hat{x} - y\hat{y})\right] \\
&= \frac{B_0}{a}(\hat{x} \times \hat{x} - \hat{y} \times \hat{y}) = 0.
\end{aligned}$$

(b) The magnetic field lines are given by the differential equation

$$\frac{dy}{dx} = \frac{B_y}{B_x} = -\frac{y}{x},$$

i.e.,

$$xdy + ydx = 0,$$

or

$$d(xy) = 0.$$

Hence

$$xy = \text{const.}$$

The field lines are shown in Fig. 5.17. In order to create such a field, four infinitely long straight currents parallel to the z direction are symmetrically placed on the four quadrants with flow directions as shown in Fig. 5.17.
(rest ommited)

InternVL2.5

(A) Verifying Maxwell's Equations in Free Space

To verify that this field satisfies Maxwell's equations, we check Gauss's law for magnetism and Ampère's law.

Step 1: Compute $\nabla \cdot \mathbf{B}$

The divergence of a magnetic field should always be zero:

$$\nabla \cdot \mathbf{B} = \frac{\partial B_x}{\partial x} + \frac{\partial B_y}{\partial y} + \frac{\partial B_z}{\partial z}.$$

For our given field,

$$B_x = \frac{B_0}{a}x, \quad B_y = -\frac{B_0}{a}y, \quad B_z = 0.$$

Thus,

$$\nabla \cdot \mathbf{B} = \frac{\partial}{\partial x}\left(\frac{B_0}{a}x\right) + \frac{\partial}{\partial y}\left(-\frac{B_0}{a}y\right) + \frac{\partial}{\partial z}(0).$$

Computing the derivatives,

$$\nabla \cdot \mathbf{B} = \frac{B_0}{a}(1) - \frac{B_0}{a}(1) = 0.$$

Since $\nabla \cdot \mathbf{B} = 0$, Gauss's law for magnetism is satisfied.

Step 2: Compute $\nabla \times \mathbf{B}$

Ampère's law states that in free space (without currents):

$$\nabla \times \mathbf{B} = \mathbf{0}.$$

To compute the curl,

$$\nabla \times \mathbf{B} = \begin{vmatrix} \hat{x} & \hat{y} & \hat{z} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ \frac{B_0}{a}x & -\frac{B_0}{a}y & 0 \end{vmatrix}$$

Expanding,

$$\nabla \times \mathbf{B} = \left(\frac{\partial(0)}{\partial y} - \frac{\partial(-B_0 y/a)}{\partial z}\right)\hat{x} + \left(\frac{\partial(B_0 x/a)}{\partial z} - \frac{\partial(0)}{\partial x}\right)\hat{y} + \left(\frac{\partial(-B_0 y/a)}{\partial x} - \frac{\partial(B_0 x/a)}{\partial y}\right)\hat{z}.$$

Since $B_z = 0$, many terms vanish:

$$\nabla \times \mathbf{B} = (0-0)\,\hat{x} + (0-0)\,\hat{y} + \left(-\frac{B_0}{a} - \left(-\frac{B_0}{a}\right)\right)\hat{z}.$$

Thus,

$$\nabla \times \mathbf{B} = 0.$$

Error: Incorrect conclusion.
The correct result is $\nabla \times \mathbf{B} = \frac{2B_0}{a}\hat{z}$, meaning there **must be a current** present. The LLM mistakenly claims no currents exist in free space.

E VISION ERROR

Questions: Switch S is thrown to position A as shown in Fig. 3.30.
(a) Find the magnitude and direction ("up" or "down" along page) of the currents in $R_1,\ R_2,$ and R_3, after the switch has been in position A for several seconds.
Solutions:
Let the currents in $R_1,\ R_2,\ R_3$ be $i_1,\ i_2,\ i_3$ respectively.
(a) When the switch is thrown to position A, we have instantaneously

$$i_1(0) = i_2(0) = \frac{2}{R_1 + R_2} = \frac{2}{3+2} = 0.4\,\text{A},$$

$$i_3(0) = 0,$$

$$i_1(\infty) = \frac{2}{R_1 + R_2 \parallel R_3} = 0.59\,\text{A}.$$

After the switch is in A for some time, we have

$$i_2(\infty) = \frac{R_3}{R_2 + R_3} i_1(\infty) = 0.12\,\text{A},$$

$$i_3(\infty) = \frac{R_2}{R_2 + R_3} i_1(\infty) = 0.47\,\text{A}.$$

As seen from the ends of L_1 the resistance in the circuit is

$$R = R_3 + R_1 \parallel R_2 = 1.7\,\Omega,$$

and the time constant is

$$\tau = L_1/R = \frac{5}{1.7} = \frac{1}{0.34}\ \text{s}.$$

Using $i(t) = i(\infty) + [i(0) - i(\infty)]e^{-t/\tau}$, (see Problem 3014), we have

$$i_1(t) = 0.59 - 0.19e^{-0.34t}\,\text{A},\ \text{the direction is upward,}$$

$$i_2(t) = 0.12 + 0.28e^{-0.34t}\,\text{A},\ \text{the direction is downward,}$$

$$i_3(t) = 0.47(1 - e^{-0.34t})\,\text{A},\ \text{the direction is downward.}$$

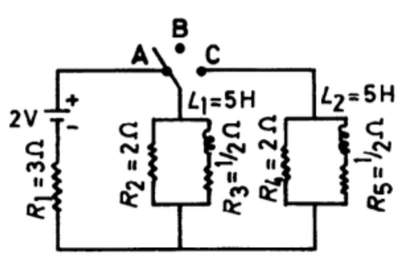


Figure 4: question image

Qwen2-VL-72B

Step (a): Steady-State Condition (Switch at Position A for Several Seconds)

Understanding the Circuit in Steady-State

omitted

Initial Currents (At $t = 0$)

omitted

Final Currents (After the Inductor Stabilizes)

- The equivalent resistance of R_2 and R_3 is:

$$R_2 \parallel R_3 = \frac{2 \times 0.5}{2 + 0.5} = \frac{1}{2.5} = 0.4\Omega \tag{19}$$

- The total resistance seen by the source:

$$R_{\text{total}} = R_1 + R_{\text{eq}} = 3 + 0.4 = 3.4\Omega \tag{20}$$

- The steady-state current through R_1:

$$i_1(\infty) = \frac{2}{3.4} = 0.59A \tag{21}$$

- Using current division:

$$i_2(\infty) = \frac{R_3}{R_2 + R_3} i_1(\infty) = \frac{0.5}{2.5} \times 0.59 = 0.12A \tag{22}$$

$$i_3(\infty) = \frac{R_2}{R_2 + R_3} i_1(\infty) = \frac{2}{2.5} \times 0.59 = 0.47A \tag{23}$$

Time Constant Calculation

- The resistance seen from the inductor:

$$R = R_3 + (R_1 \parallel R_2) = 0.5 + 1.2 = 1.7\Omega \tag{24}$$

- Time constant:

$$\tau = \frac{L_1}{R} = \frac{5}{1.7} = 2.94s \tag{25}$$

- Transient response:

$$i_1(t) = 0.59 - 0.19e^{-0.34t} \quad \text{(upward)} \tag{26}$$

$$i_2(t) = 0.12 + 0.28e^{-0.34t} \quad \text{(downward)} \tag{27}$$

$$i_3(t) = 0.47(1 - e^{-0.34t}) \quad \text{(downward)} \tag{28}$$

Mistake in the solution

The image-reading errors stemmed from misinterpreting circuit connections and misreading component values. The LLM incorrectly treated R_2 as series with R_1 instead of parallel with R_3, overestimating total resistance. It also misread R_3 as 3Ω instead of 0.5Ω, distorting current calculations. Additionally, the time constant was miscalculated by using $R_2 + R_3$ instead of $R_3 + (R_1 \parallel R_2)$, leading to incorrect transient responses. These errors affected both steady-state and transient analyses, highlighting the need for careful circuit interpretation and accurate component identification.