Evaluating Large Language Models in Olympic-Level Physics Problems: A Benchmark Dataset

Anonymous ACL submission

Abstract

Large Language Models (LLMs) and Large Multimodal Models (LMMs) have demonstrated notable capabilities across a wide range of tasks and domains, showcasing advanced problem-solving skills that encompass everything from natural language understanding and generation to complex decision-making processes. However, the extent of their proficiency in tackling mathematical physics problems remains relatively underexplored. In this paper, we propose PhOPile, a high-quality, multimodal, physics-specific, and Olympic-level physics dataset. We detail the meticulous process of data collection, cleaning, and structuring to ensure the dataset's integrity and utility. Furthermore, we conduct a high-granularity evaluation of the performance of currently popular LLMs and LMMs on our dataset and provide a benchmark of their physics problem-solving capability and enrich assessment options for models' competencies in natural subjects. We also introduce an evaluation method that enables a more detailed measurement of the model's reasoning capabilities. Our research represents the first attempt to reveal the potential and current limitations in interpreting and solving complex physics challenges, setting a foundational baseline for subsequent advancements in this field.

1 Introduction

Large language models (LLMs) and large multimodal models (LMMs) such as GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023), and Gemini (Deepmind, 2023) have shown exceptional performance in tasks beyond Natural Language Processing (NLP). Not surprisingly, there is an increasing research interest in exploring scientific domains, especially in mathematics. Significant progress has been shown such as MathPrompter (Imani et al., 2023) using prompt template, FunSearch (Romera-Paredes et al., 2024) utilizing self-iteration, Improving LLM Fine-Tuning method

Figure 1: Source of PhOPile

(Liu et al., 2023), DeepSeekMath (Shao et al., 2024) leveraging improved reinforcement learning, Llemma (Azerbayev et al., 2023) and Goat (Liu and Low, 2023). Typically, these models need training to specifically address mathematical problems, which often necessitates a substantial quantity of high-quality datasets or corpus like MiniF2F (Zheng et al., 2021), MATH (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), MLFMF (Bauer et al., 2023), Mathpile (Wang et al., 2023b) and Proof-Pile-2 (Azerbayev et al., 2023).

However, research focusing on LLMs and LMMs of physic reasoning remains notably sparse. There is no physics specified model at present. Related studies mainly focus on natural science datasets and corpora used for LLMs and LMMs training, including a portion of low-difficulty physics data such as SciQ (Johannes Welbl, 2017), ScienceQA (Lu et al., 2022) and TheoremQA (Chen et al., 2023). Nevertheless, physics, as an exceedingly important discipline within the natural sciences, plays a pivotal role in various fields such as construction, aerospace, and electronic engineering. It is still insufficient to evaluate the physical inferencing abilities of LLMs with the

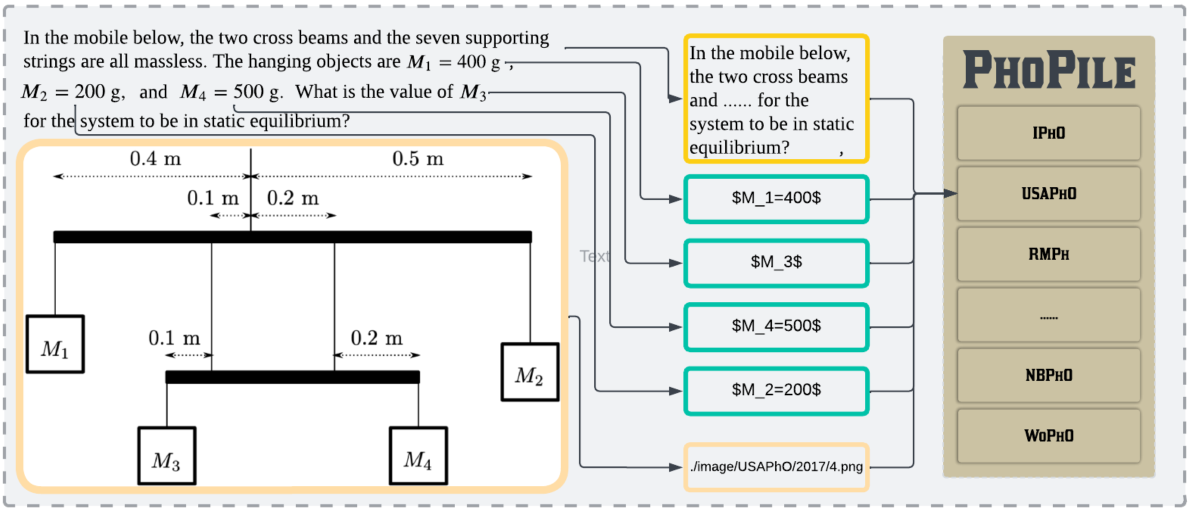


Figure 2: PhOPile collection and processing process. We collect it from physics competitions around the world. It has gone through a series of preprocessing, including: regular character collection, mathematical formula conversion, image extraction, question number modification, etc. Detailed steps are clearly stated in Section 3.

current datasets while they are relatively small, not difficult enough, and not covering a complete range of physical knowledge.

By recognizing the lack, in this paper, we introduce PhOPile, a multimodal dataset featuring Olympiad-difficulty-level physics problems. Specifically designed for fine-tuning and evaluating LLMs' and LMM's capabilities in solving advanced physics problems, PhOPile marks a pioneering step in bridging the gap between artificial intelligence and high-level physics problem-solving. In contrast to the rigorous logic demanded in mathematical problem-solving datasets, physics problems call for a distinct form of reasoning that often involves the integration of a broad spectrum of conceptual understanding (Hung and Jonassen, 2006) along with mathematical analysis. Acknowledging this difference, we utilize PhOPile, which comprises physics competition problems, to assess the efficacy of LLMs and LMMs. Our goal is to share our findings with the community, thereby fostering further advancements in this field by fellow scholars.

In summary, our contributions are summarized as follows:

- We introduce a carefully processed, extremely high-quality, and first of the kind physics problems dataset of physics problems with solution steps and their relevant images.
- We provide a new metric of physics problems for evaluating LLMs.
- We evaluate this dataset among the current mainstream LLMs and LMMs and propose a benchmark, which not only clarifies their abilities in physics but also identifies opportunities for enhancement.
- We discuss the potential future developments of LLMs in the field of natural sciences and their implications for related domains.

2 Data Collection

We collect of Physics Olympiad questions from various regions around the globe as shown in Fig. 1, including: International Physics Olympiad (IPhO, 1967-2023), Asian Physics Olympiad (APhO, 2000-2021), European Physics Olympiad (EuPhO, 2017-2021), Nordic-Baltic Physics Olympiad (NBPhO, 2003-2021), Romanian Master of Physics (RMPhO, 2012-2021), United States Physics Olympiad (AAPT, 2007-2019), and British Physics Olympiad (BPhO, 2001-2022).

Physics problems often involve mathematical formulas, we use LaTeX to formulate the solutions with all the detailed information in plain text. A representative sample question (from USAPhO 2017) from our dataset is shown in Fig. 2. To enhance our work efficiency and ensure the uniformity and appropriateness of our dataset in LaTeX format, we utilize MathPix, an OCR recognition software, to convert the content of images into LaTeX code. For finalized dataset samples stored in JSON format, please refer to the Appendix A.1. The following section will introduce the steps we took in data processing.

3 Data Processing

A physics question problem has multiple parts: text, numbers, and images. Fig. 2 illustrates how we process the each part of a physics question. The following section will introduce detailed steps we take in data processing.

Sub-questions Unlike the most mathematical datasets, a considerable number of physics competition questions consist of multiple sub-questions that are interconnected. We collect and specify these questions by using Arabic number as their index. An completed example is shown in Fig. A.1.

Hint Information & Solution Format Requirements Physics competition problems sometimes provide additional information, which, in the sources we compile, are referred to as: 'Hint', 'Data', or 'For information only'. Such information usually contains the premises necessary to solve the problem. Therefore, depending the position appears, we place this type of information after the corresponding question text to ensure a more targeted and effective use, enhancing the specificity and relevance of the information provided in relation to that specific part of the problem.

For questions that specify formatting requirements for the candidate responses, for example, 'please round to two decimal places', the positioning of such instructional text is aligned with the previously mentioned approach for hint information.

Images in the Question A significant portion of the questions and answers in our dataset include one or more images. For such questions, we have stored the related images in a folder and introduced '`imgQ`' and '`imgA`' in our dataset, which stores the local URLs for the images associated with the questions and the solutions, respectively.

At the same time, in order to record the position where the image appears in the question text, we replaced words like 'figure' with a mark: `###img_N###`, where 'N' denotes the order of the image. An example can be observed in Fig. 7b. The '`###img_1###`' in '`Question`' and '`Solution`' stands for the image of '`./pic/question/34.png`' and '`./pic/answer/65.png`' respectively in repository.

Regarding the captions of these images, we observe two primary types. The first type consists merely of image labels, such as 'Fig. 05' or 'figure 1'. For these cases, we choose to omit the captions from our dataset, as they provide minimal informational value. The second type of captions such as 'Figure 1: Isosceles glass prism with an apex angle of 90°', containing crucial content of the problem are added in question text. Similar method has been proven to improve the accuracy of language model outputs by MathVista (Lu et al., 2024), as they input with both the captions of images and text from enhanced Optical Character Recognition (OCR).

Data cleaning and filtering We delete extraneous elements from the questions, including historical background introductions, scoring criteria, and regulations or policies related to competitions. We filtered out certain LaTeX commands solely involved in adjusting the format, as they contribute nothing to the essence of the question. This serves the dual purpose of noise reduction and data compactness. Additionally, this approach aims to meet the text length requirements specified for the window length of LLMs.

Multiple Solutions For certain questions, the source files provide multiple solutions. They often appear as 'Solution 2', 'Another way to solve this problem'. To adhere to the specificity of training and ensure the constraints of text length within the window parameters, we establish a new key called '`solution2`' or '`solution3`' (if there are more alternative solutions) to store them.

4 Data Analysis

Summary We collect 4,496 high-quality, and Olympic-level physics competition questions from 1967 to 2021 in various fields. The specific amount of each source is shown in Fig. 3 and more detailed token statistics from different sources are shown in Table 1.

At the present stage, the advancement in research pertaining to LMMs addressing problems involving images remains suboptimal. Notably, some images can be considered meaningless, as shown in Fig. 5 (b), images for this type of topic do not contain any useful information, whereas images Fig. 5 (a) contain necessary information for solving the question. Therefore, questions containing images like Fig. 5 (b) can be regarded as a complete question

Source	# Questions	# Tokens	Max (#Tokens)	Min (#Tokens)	Ave (#Tokens)	Years
APhO	589	279,802	3,208	18	475	2000-2021
EuPhO	23	20,738	3,951	20	902	2017-2021
IPhO	947	446,170	4,487	21	471	1967-2021
NBPhO	429	137,406	1,638	17	320	2003-2021
RMPh	195	81,364	1,861	49	420	2012-2021
USAPhO	723	136,179	2,030	11	188	2007-2019
WohO	146	66,038	1,737	19	452	2011-2013
Total	4496	1,275,406	-	-	-	-

Table 1: The token statistics of each components of PhOPile.

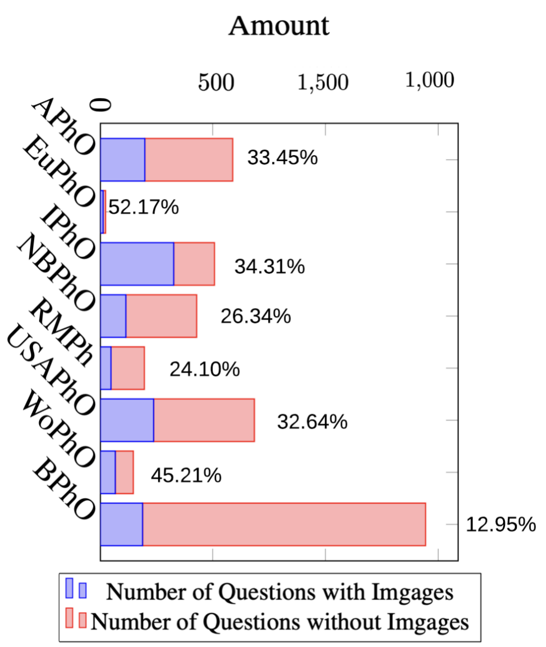


Figure 3: Comparison of Number of Questions from Different Sources with and without Images. The percentage represents the proportion of questions with images.

even without images. Models can still obtain all information about the question without images in the prompt.

The Token Count Distribution of Answers and Solutions The token count statistics are illustrated in Fig. 4. Due to the excessive length of certain individual questions (for example, some questions spend huge spaces to explain a complex and comprehensive physical environment\premises or some solutions are subjective, and they often use a lot of tokens to explain a physical phenomenon clearly). We can see that the number of tokens in most questions and solutions is within 500. For the current popular LLMs and LMMs, the window for text length allows for a complete round of training without the need to prune the prompt.

Ambiguous Statistical Values Explanation The lowest number of token is 1, typically representing sub-questions that serve as a completion of their preceding counterparts. For example, the question '`Please determine the average speed of the period (i) T and (ii) 2T.`' has two sub-questions: The one is '`Please determine the average speed of the period T`' with the token count is 9 and the other one is '`2T.`' with the token count is 1. The lowest number of solution tokens could be 0, because some questions require candidates to sketch a graph, so there is no literal solutions.

Images in Solutions In our dataset, a great part of solutions incorporate images. Some of these images are curve graphs, designed to illustrate the relationship between two variables as specified by the question, while others are structural diagrams for force analysis, among other types. However, current LMMs are incapable of generating images in accordance to the specific requirements posed by the questions. Nevertheless, we have included these instances in PhoPile, aiming to facilitate future research contributions by scholars.

4.1 Data Contamination

We conducted data contamination detection on PhoPile. Due to the limited current research on physical sciences by both LLMs and LMMs, we were unable to find a perfectly corresponding benchmark dataset. However, we discovered some natural science datasets that include physics problems: SciQ (Johannes Welbl, 2017), ScienceQA (Lu et al., 2022) and TheoremQA (Chen et al., 2023). We performed n-grams data contamination

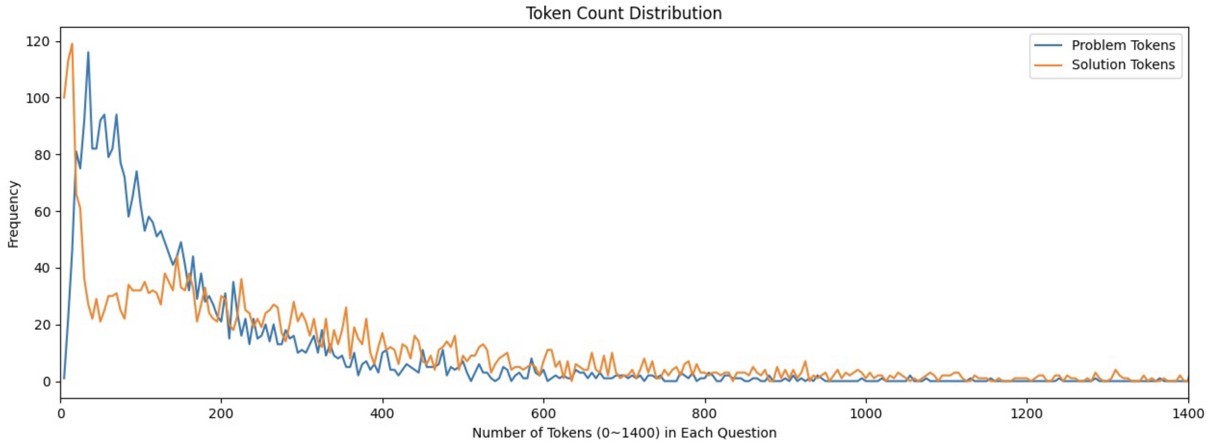


Figure 4: Token Count Distributions in Problems and Solutions: The images displays the token distribution statistics of PhoPile, while blue line and yellow line stand for the token distribution statistics of the problem text and solution text respectively. Please note that this picture only shows the distribution of tokens between 0 and 1400, since few questions with exceptionally long tokens (over 4000) were excluded from the statistical image to avoid compression that would render the image unclear

monitoring on them, setting n to 3. The results were as follows: SciQ: 4.2977%, ScienceQA: 0.7759%, and TheoremQA: 10.0088%. Above results, even when setting n as low as 3, indicate remarkably low contamination rates. These figures suggest that the quality of our dataset is exceptionally high, which is positive for our research integrity and the validity of the data we utilize. However, it's important to acknowledge the limitations in guaranteeing the absence of contamination in our dataset in relation to the training sets of closed-source models. Since the datasets used for training these models are not publicly available, we cannot definitively ensure that our data is free from overlap with these their training data.

5 Experiments

In this section, we use the test set and perform evaluations on it using LLMs and LMMs.

5.1 Test Dataset

Due to the window context length limit of Llama2 (Touvron et al., 2023) is 4096 and the average solution token length in PhoPile being around 400, we initially filtered out questions whose total number of tokens exceeded 3500 in our test set. By doing so, we can ensure that the prompts input into the model are within the window text length range of the model. Then, we extracted two sets of questions for testing, employing random selection to ensure equal distribution across sources and by year. Finally, we constructed *Test-PhoPile* and *Test-PhoPile-V*, consisting of 50 main questions with 82 sub-questions in pure text, and 20 main questions with 73 sub-questions including image queries, respectively.

5.2 Evaluation

To enable models to provide highly targeted responds to each sub-question of problems containing multiple sub-questions, we do not simply concatenate the text of all sub-questions and input them into models. Instead, our prompts strategy is: We input the sub-question texts in the order and obtain the corresponding response text. So the message history is a sub-question text sent by the user, and then a respond from the assistant arranged in sequence. An prompt example of the format of input for a text-based question is demonstrated in Fig. 19.

Physics questions like the International Olympiad often require candidates to answer questions with long logical chains of reasoning. Hence, the scoring criteria should not solely assess the provision of the final answer. Whether solving it partially correctly or entirely incorrectly from the outset, neither leads to the correct final answer. However, we posit that the LLMs that achieves partial correctness demonstrates a superior ability in handling physics questions. Consequently, we have opted for a step-by-step scoring strategy.

In the original examination marking scheme, each question was assigned varying points, a sys-

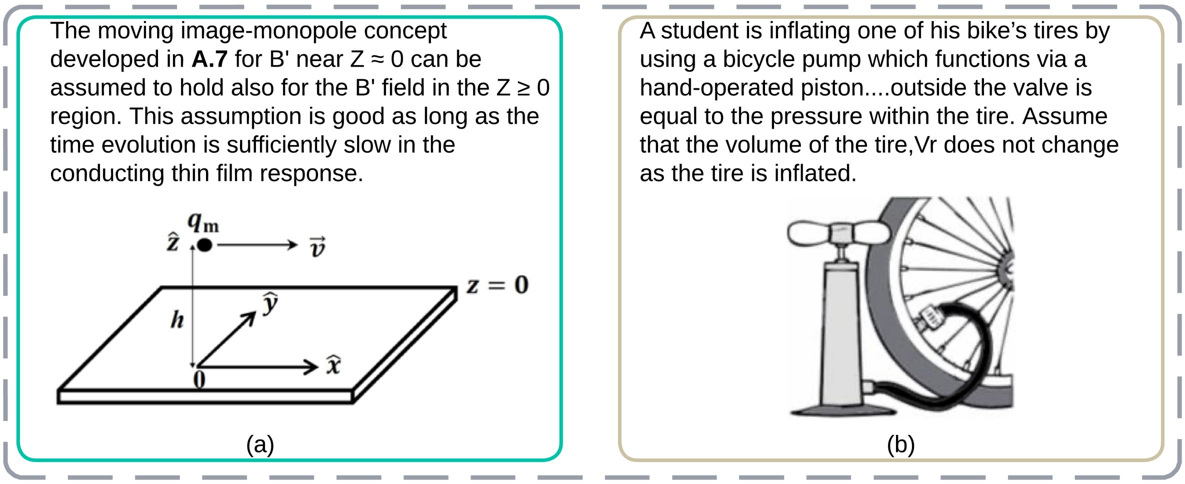


Figure 5: Examples of Questions with Meaningless and Meaningful Images in PhoPile: Part (a) is the question with its image providing necessary condition. Image of question demonstrated in Part (b) does not show any useful information in problem solving.

Model	Input	FCR (%)	FIR (%)	PCR (%)	APC (%)	ALL (%)
Llama2-13B	*T*	1.20	71.08	27.71	28.70	9.16
ChatGPT-3.5	*T*	14.63	15.85	69.51	40.18	42.56
Gemini-Pro	*T*	15.85	18.29	65.85	41.48	43.17
ChatGPT-4	*T*	35.37	3.66	60.98	59.00	71.34
Zero-Shot Chain of Thought						
CoT Llama2-13B	*T*	1.22	71.95	26.83	29.55	9.15
CoT ChatGPT-3.5	*T*	12.20	13.41	74.39	12.50	45.49
CoT Gemini-Pro	*T*	9.76	17.07	73.17	35.67	35.85
CoT ChatGPT-4	*T*	40.24	6.10	53.66	56.36	70.49
Large Multimodal Models (on *Test-PhoPile-V*)						
Gemini-Pro-Vision	*T, I*	4.11	23.29	72.60	35.28	29.73
ChatGPT-4-Vision	*T, I*	17.81	6.85	75.34	48.55	54.38

Table 2: The evaluation results of currently popular large language models under different input methods on PhOPile. Input: *Text*: Question text only. *Text, Image*: Question text and images. To further increase consistency, the score for each question is the mode of the scores among 10 times.

tem that proved challenging and inconsistent for grading purposes. Consequently, we adopted a new approach, each initial order question is uniformly valued at ten points. A full score is awarded when candidate LLM generates the correct answer. In cases where the LLM's response is incomplete or partially correct, the score is assigned based on the extent of correctness (the furthest correct element in the reasoning process) provided by the LLM. This modification aims to standardize scoring while accommodating the varying completion levels of solutions generated by the LLM. We leverage ChatGPT-4 to grade every question by letting it compare the standard solution and the solution output by candidate LLM. Then, we compute the arithmetic mean of the scores obtained for all questions answered by candidate LLMs and LMMs. For examples of prompting ChatGPT-4 to mark a solution, please refer to the Appendix 20.

5.3 Validation of GPT-4 Scoring Accuracy

We develop a new marking method which can improve the accuracy of evaluation for reasoning chain solutions without human intervention by inputting the standard solution and the solution generated by model. To ensure consistency, we conducted a series of experiments before using GPT-4 to score LLMs' and LMMs' results. These experiments include, but are not limited to, combinations of the following situations:

- Provide a generated fully correct solution. (GPT-4 gives 10)
- Provide half of the generated fully correct solution. (GPT-4 gives 1 to 9)
- Provide a completely wrong solution that need to be scored. (GPT-4 gives 0)
- Slightly change the final answer but the error is within the reasonable range required by the question. (GPT-4 gives 10)
- Provide answers in normal format, Latex format, and answers without units but the numbers are the same. (GPT-4 gives 10)
- Provide a generated fully correct solution with changing the final answer to a totally wrong one. (GPT-4 gives 9)
- Provide a fully incorrect solution but with the correct final answer. (GPT-4 gives 0)

All of the above are situations where GPT-4 can correctly respond to a candidate's answers. However, for some answers, GPT-4 may also give scores that do not meet the scoring criteria.

- Slightly modify the numerical value of the final answer, but do not enter a reasonable range. For example, the assessed answer is 3.8, whereas the standard answer is 4. (GPT-4 gives 9; Should 0)
- Cut a generated fully correct solution to a half and add the correct final answer. (GPT-4 gives 10; Should 1 to 9)
- Randomly delete some scoring points from a generated fully correct, but the final answer is correct. (GPT-4 gives 10)

Detailed sample questions, answers, and prompts of this experiment are shown in Appendix A.2.

Above three situations do not align with the marking criteria. But the first one can be easily avoid via prompt. For the second and the third, according to the scoring rules, an answer that is missing some scoring points should not receive full marks, which is different from the results provided by GPT-4. However, such cases are very rare. LLMs sometimes do exhibit leaps in logic that result in missing scoring points. But what we are measuring here is the reasoning capability of these models. Even in cases where there are leaps in logic, this does not negate their ability to reason, as they are capable of producing the final result. The results of these experiments show that our method can give correct judgments when facing various levels and types of answers to be scored.

Evaluation metrics:

- Full Correctness Rate (FCR): The percentage of the number for completely correct questions to the total number of questions.
- Full Incorrectness Rate (FIR): The percentage of the number for completely incorrect questions to the total number of questions.
- Partial Correctness Rate (PCR): The percentage of the number for partially correct questions to the total number of questions. Please note that FCR+FIR+PCR=100%.
- Average score of Partially Correct questions (APC): The arithmetic mean of all answers that are partially correct.
- ALL: Overall correctness rate, the arithmetic mean of all questions scores.

5.4 Results

The evaluation results can be seen in Table 2. Overall, ChatGPT-4 consistently achieved the best performance under all conditions, followed by Gemini Pro and ChatGPT-3.5, with the untrained open-source model Llama2-13B at last. Specifically, in the context of Zero-Shot CoT reasoning, ChatGPT-4 exhibited a 5% increase in FCR. However, it is noteworthy that empirical observations revealed a varying degree of performance degradation for Gemini Pro and ChatGPT-4 in the context of Zero-Shot CoT reasoning, with Gemini experiencing a decline of 7.32 percentage points.

For LMMs results, the comprehensive performance of ChatGPT-4 surpasses that of Gemini by nearly double. However, according to the results from MathVista (Lu et al., 2024), in the domain of mathematical imagery, the capability of ChatGPT-4-V is remarkably similar to Bard (AI, 2023) which is the precursor to Gemini. The primary reasons for this significant discrepancy are attributed to the inherent differences in the disciplines, the nature of the problem images, and, secondarily, the differences in the evaluation methodologies. An analysis of Gemini's outputs reveals that, often, Gemini struggles to comprehend the inputted physical images, with examples of Gemini's outputs available in the Appendix A.5.

6 Related Work

Study of LLMs in Natural Science Subjects. In recent years, LLMs have developed very rapidly, providing great convenience for people's needs in all aspects of life. These models, like GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023) and Gemini (Deepmind, 2023) have already shown great performance in terms of accuracy, interpretability, and multimodality, similarly as general LLMs, they show outstandingly high performance of natural science Q\A and mathematical reasoning. Meanwhile, a range of excellent open source models, including T5 (Raffel et al., 2020), ChatGPT-2 (Radford et al., 2019) and Llama2 (Touvron et al., 2023), is available for researchers to enhance further, by training them on a specialized dataset to attain superior capabilities compared to generalized models. Consequently, a series of outstanding open source models that are specifically trained and fine-tuned on math have emerged, such as DeepSeek-Math (Shao et al., 2024), Llema (Azerbayev et al., 2023) and Goat (Liu and Low, 2023). Additionally, there are also a few models focusing on formal proof such as LeanDojo (Yang et al., 2023) and LEGO-Prover (Wang et al., 2023a); these are models trained on math-specialized corpus or datasets. However, in the expansive domain of mathematics, the multitude of sub-disciplines presents a significant challenge for models with constrained parameters to adequately address comprehensive mathematical problems. Studies like Boosting LLM Reasoning (Huang et al., 2023a) and LeanDojo (Yang et al., 2023) use a retrieval-augmented approach to improve the accuracy of mathematical problem-solving. It is noteworthy that research at the intersection of linguistics and natural sciences remains relatively scarce. Scholars have placed a greater emphasis on mathematical reasoning.

Natural Science datasets for LLMs. Models which demonstrate excellent performance on mathematical ability are inseparable from high-quality datasets and corpus such as Mathpile (Wang et al., 2023b), proof-pile-2 (Azerbayev et al., 2023), MiniF2F (Zheng et al., 2021), MATH (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), MLFMF (Bauer et al., 2023)) and the corpus proposed by DeepSeekMath (Shao et al., 2024). The aforementioned datasets consist solely of textual data; however, it is commonly understood that the interpretation of mathematical problems often requires the analysis of images. Consequently, MathVista (Lu et al., 2024) introduced a specialized image-based mathematical dataset and conducted evaluations of models such as GPT-4 (OpenAI, 2023), ChatGPT (OpenAI, 2022), Claude-2 (Anthropic, 2023), and mPLUG-Owl-LLaMA (Ye et al., 2023) from various perspectives: purely textual input, text with captions and image OCR (Augmented-LLMs), and multimodal analysis. However, there is a noticeable paucity of specialized research linking LLMs with the discipline of physics. The relevant work in this area is confined to a minimal subset of physics-related data within certain natural science datasets, such as SciQ Dataset (Johannes Welbl, 2017), ScienceQA (Lu et al., 2022), C-eval (Huang et al., 2023b), E-EVAL (Hou et al., 2024), and TheoremQA (Chen et al., 2023).

7 Conclusion

In this work, we introduce PhOPile, a benchmark constructed to comprehensively test LLMs' abilities in physics reasoning under both pure text questions and image-based questions. We evaluated four mainstream models, among which ChatGPT-4 showed the best performance. Furthermore, we proposed a new marking method tailored for problems that contain multiple sub-questions and involve reasoning steps. This method allows for a more sophisticated evaluation of language models' reasoning capabilities, providing deeper insights into their ability to process and solve complex tasks that require step-by-step logical deduction.

8 Limitations

Due to the intrinsic characteristics of physics problems, which often comprise numerous sub-questions and exceptionally lengthy prompts, certain measures must be adopted by language models constrained by a shorter window text length to facilitate comprehensive problem fine-tuning and evaluation. In this work, to ensure the progress and to control variables of the evaluation, only those problems with a total token count of less than 3500 were selected for testing. In addition, we cannot detect data pollution for the closed-source models of OpenAI and Deepmind. Finally, for questions that require drawings in the answer, all existing models cannot provide effective answers.

References

AAPT. 2007-2019. United states physics olympiad official website. Accessed: 2024-02-15.

Google AI. 2023. Bard: A large language model.

Anthropic. 2023. Claude 2: Improvements and capabilities of anthropic's second-generation ai chatbot. `https://www.anthropic.com/news/claude-2`. Accessed: 2024-02-14.

APhO. 2000-2021. Asian physics olympiad 2019 official website. Accessed: 2024-02-15.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2023. Llemma: An open language model for mathematics.

Andrej Bauer, Matej Petković, and Ljupčo Todorovski. 2023. Mlfmf: Data sets for machine learning for mathematical formalization. *https://arxiv.org/abs/2310.16005*.

BPhO. 2001-2022. British physics olympiad official website. Accessed: 2024-02-15.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Wenhu Chen, Ming Yin, Max Ku, Elaine Wan, Xueguang Ma, Jianyu Xu, Tony Xia, Xinyi Wang, and Pan Lu. 2023. Theoremqa: A theorem-driven question answering dataset. *arXiv preprint arXiv:2305.12524*.

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*.

Deepmind. 2023. Gemini: A family of highly capable multimodal models. *ArXiv*, abs/2312.11805.

EuPhO. 2017-2021. European physics olympiad official website. Accessed: 2024-02-15.

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. *NeurIPS*.

Jinchang Hou, Chang Ao, Haihong Wu, Xiangtao Kong, Zhigang Zheng, Daijia Tang, Chengming Li, Xiping Hu, Ruifeng Xu, Shiwen Ni, and Min Yang. 2024. E-eval: A comprehensive chinese k-12 education evaluation benchmark for large language models. *ArXiv*, abs/2401.15927.

Xijie Huang, Li Lyna Zhang, Kwang-Ting Cheng, and Mao Yang. 2023a. Boosting llm reasoning: Push the limits of few-shot learning with reinforced in-context pruning. *ArXiv*, abs/2312.08901.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Fanchao Qi, Yao Fu, Maosong Sun, and Junxian He. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *ArXiv*, abs/2305.08322.

Woei Hung and David H. Jonassen. 2006. Conceptual understanding of causal reasoning in physics. *International Journal of Science Education*, 28(13):1601–1621.

Shima Imani, Liang Du, and H. Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. In *Annual Meeting of the Association for Computational Linguistics*.

IPhO. 1967-2023. International physics olympiad unofficial website. Accessed: 2024-02-15.

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

Tiedong Liu and Kian Hsiang Low. 2023. Goat: Fine-tuned llama outperforms gpt-4 on arithmetic tasks. *ArXiv*, abs/2305.14201.

Yixin Liu, Avi Singh, C. Daniel Freeman, John D. Co-Reyes, and Peter J. Liu. 2023. Improving large language model fine-tuning for solving math problems. *ArXiv*, abs/2310.10047.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

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

NBPhO. 2003-2021. Nordic-baltic physics olympiad official website. Accessed: 2024-02-15.

OpenAI. 2022. Chatgpt. `https://openai.com/blog/chatgpt`. Accessed: 2023-02-06.

OpenAI. 2023. Gpt-4: Openai's generative pre-trained transformer 4 model.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

RMPhO. 2012-2021. Romanian master of physics 2023. Accessed: 2024-02-15.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Andrey Novikov, et al. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, v1.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Haiming Wang, Huajian Xin, Chuanyang Zheng, Lin Li, Zhengying Liu, Qingxing Cao, Yinya Huang, Jing Xiong, Han Shi, Enze Xie, Jian Yin, Zhenguo Li, Heng Liao, and Xiaodan Liang. 2023a. Lego-prover: Neural theorem proving with growing libraries.

Zengzhi Wang, Rui Xia, and Liu Pengfei. 2023b. Generative ai for math: Part i – mathpile: A billion-token-scale pretraining corpus for math. *arXiv preprint arXiv:2312.17120*.

Kaiyu Yang, Aidan Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. 2023. LeanDojo: Theorem proving with retrieval-augmented language models. In *Neural Information Processing Systems (NeurIPS)*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yi Zhou, Junyan Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qiang Qi, Ji Zhang, and Feiyan Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *ArXiv*, abs/2304.14178.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. 2021. Minif2f: a cross-system benchmark for formal olympiad-level mathematics. *arXiv preprint arXiv:2109.00110*.

A Appendix

A.1 Question Examples in PhOPile

To fully record the details of the questions, we create 'question_number', 'sub_question_number', and 'sub_sub_question_number' which stand for question number, first order sub-question number and second order sub-question number in our dataset to facilitate distinction. Furthermore, within the text of the questions, we replace the original question index, which typically consist of Arabic numerals, English letters, Roman numerals, etc, with Arabic number in '`sub_question_number`' or '`sub_sub_question_number`'. Fig. 7 shows an example of how we process a question with second order questions.

a) A steel ball is thrown down with a speed of $3.0\,\mathrm{m\,s^{-1}}$ on to a hard surface from a height of $2.0\,\mathrm{m}$. It retains 70% of its energy on each bounce. Calculate

(i) the speed at which it hits the ground for the first time, and

(ii) the maximum height it reaches after the 4th bounce.

[2]

Figure 6: Raw Question in BPhO

```
Question: A steel ball is thrown down with a speed of $3.0 \mathrm{~m} \mathrm{~s
    }^{-1}$ on to a hard surface from a height of $2.0 \mathrm{~m}$. It retains $70
    \%$ of its energy on each bounce. Calculate the speed at which it hits the
    ground for the first time
solution: $\begin{aligned} & v^2=u^2+2 a s \\ = & 3^2+2 \times 9.81 \times 2 \\ = &
    6.946=7.0 \mathrm{~ms}^{-1}\end{aligned}$
question number: 1
sub question number: 1
Question image path: null
Answer image path: null

Question: the ma×imum height it reaches after the 4th bounce.
Solution: $\begin{gathered}g h_1=\frac{1}{2} v^2 \eta=\frac{1}{2} \times 6.95^2 \
    times 0.7 \rightarrow h_1=1.72 \mathrm{~m} \\ h_n=h_1 \times \eta^{n-1}=0.59 \
    mathrm{~m}\end{gathered}$
question_number: 2
Question image path: null
Answer image path: null
```

(a) Demonstration of Question Numbers Processing: We replace sub-question number '(i)' and '(ii) to Arabic numeral '1' and '2'. As there is no image provided in the question, we set 'null' in image key.

```
Question: Consider a plane-parallel transparent plate, where the refractive inde×,
    $n$, varies with distance, $z$, from the lower surface (see ###img_1###). Show
    that $n_A \sin \alpha=n_B \sin \beta$. The notation is that of the figure.
Solution: From the ###img_1### we get $n_A \sin \alpha=n_1 \sin \alpha_1=n_2 \sin \
    alpha_2=\ldots=n_B \sin \beta$
Question image path: ./pic/question/34.png,
Answer image path: ./pic/answer/65.png
```

(b) Demonstration of Image Processing: We replace 'figure 1' in text to special marks to record the position of where image appears. At the same time, we input the image path to 'image' key.

Figure 7: Examples from the dataset.

A.2 Prompts and Results for Validation of GPT-4 Scoring Accuracy Experiment

Provide complete problem-solving steps that need to be scored and standard answers. (GPT-4 gives 10) [Proof question]

Input 1:

System:
You are a professional physicist and you will grade answers provided by physics students by reference to standard answers. The full score is 10 points, and the minimum score is 0 points. If the student gives the final answer, full marks will be awarded directly. If the student does not give the final answer or the final answer is incorrect, please score based on the proportion of correct calculation steps given by the student. You only need to output a score number.
User:
Standard answer:
If the light falls at the angle of $30^{\\circ}$ then the film seen from the perpendicular direction cannot be colored. It would appear dark.
Student answer:
The color seen in a soap film is due to constructive and destructive interference patterns of light reflecting off the front and back surfaces of the film. The thinnest possible soap film that still displays a color will be when the film creates the condition for constructive interference for the shortest visible wavelength of light (violet) in the first order of interference (where \\(m=1\\)), observed perpendicularly (\\(\\alpha = 0\\), meaning \\(\\beta = 0\\) and thus \\(\\cos(\\beta) = 1\\)).\n\n### Constructive Interference Condition\n\nFor constructive interference from a thin film observed perpendicularly (\\(\\alpha=0^{\\circ}\\) and hence \\(\\beta=0^{\\circ}\\)), the condition for the \\(m\\)-th order of maxima is:\n\n\\[\\delta = 2d \\cdot n = m\\lambda, \\quad m=1,2,3...\\]\n\nWhere:\n- \\(\\delta\\) is the path difference that causes constructive interference.\n- \\(d\\) is the thickness of the film.\n- \\(n\\) is the refractive index of the film.\n- \\(\\lambda\\) is the wavelength of light in the film.\n- \\(m\\) is the order of interference.\n\nFor the thinnest possible film displaying color, \\(m=1\\) (first order), and we aim to solve for \\(d\\) (thickness).\n\n### Wavelength of Light in the Film\n\nThe wavelength of light in the film (\\(\\lambda\\)) is related to the wavelength in the vacuum (\\(\\lambda_0\\)) by:\n\n\\[\\lambda = \\frac{\\lambda_0}{n}\\]\n\nSince we are interested in visible light and the thinnest film would theoretically reflect violet light (since violet has the shortest wavelength in the visible spectrum, around \\(400 \\, \\mathrm{nm}\\)), we would use \\(\\lambda_0 \\approx 400 \\, \\mathrm{nm}\\).\n\n### Solving for the Minimum Thickness \\(d\\)\n\nSet \\(m=1\\) for the first order of maximum (bright fringe) and solve for \\(d\\) for violet light to find the thinnest film:\n\n\\[2d \\cdot n = \\lambda\\]\n\n\\[2d = \\frac{\\lambda_0}{n}\\]\n\nGiven that \\(\\lambda_0 \\approx 400 \\, \\mathrm{nm}\\) for violet light and substituting, we get:\n\n\\[d = \\frac{400 \\times 10^{-9} \\, \\mathrm{m}}{2n}\\]\n\nBecause the refractive index \\(n = 1.33\\) (for the soap film, similar to water), we substitute \\(n\\) into the equation:\n\n\\[d = \\frac{400 \\times 10^{-9} \\, \\mathrm{m}}{2 \\times 1.33}\\]\n\n\\[d \\approx 150.38 \\times 10^{-9} \\, \\mathrm{m}\\]\n\n\\[d \\approx 150 \\, \\mathrm{nm}\\]\n\n### Conclusion\n\nThe thinnest possible soap film that still displays a color when viewed perpendicularly would reflect violet light, characteristic of the first order of constructive interference, with a minimum thickness of approximately \\(150 \\, \\mathrm{nm}\\). This result follows from the constructive interference condition, considering the smallest possible constructive interference for visible light occurs with violet light due to its shortest wavelength in the visible spectrum.

Output 1:

Assistant: 10

Figure 8: Prompt Example of Providing Complete Problem-Solving Steps That Need to Be Scored and Standard Answers. (Proof Question)

Provide complete problem-solving steps that need to be scored and standard answers. (GPT-4 gives 1-9) [Calculation Question]

Input 1:

System:
You are a professional physicist and you will grade answers provided by physics students by reference to standard answers. The full score is 10 points, and the minimum score is 0 points. If the student gives the final answer, full marks will be awarded directly. If the student does not give the final answer or the final answer is incorrect, please score based on the proportion of correct calculation steps given by the student. You only need to output a score number.
User:
Standard answer:
Energy (3) of the question sheet can be interpreted as describing a system of a mass $\\mu$ in a circular orbit with angular velocity $\\Omega$, radius L, around a mass M (at rest). Equating the gravitational acceleration to the centripetal acceleration:$G \\frac{M}{L^2}=\\Omega^2 L$. This is indeed Kepler's third law (for circular orbits). Then, from (7), $E=-\\frac{1}{2} G \\frac{M \\mu}{L}$.**$\\beta=-\\frac{1}{2}$**.
Student answer:
To simplify the total energy expression and arrive at a formula such as $E = \\beta G \\frac{M \\mu}{L}$, we need to incorporate both kinetic and potential energy terms and their relation in a two-body system of circular orbits. \n\nFrom previous considerations, the total energy E of the system combines both kinetic (K) and potential (U) energies:\n\n\\[E = K + U\\]\n\nFrom the provided equation and context, the potential energy of two bodies attracting each other gravitationally is:\n\n\\[U = -G \\frac{M \\mu}{L}\\]\n\nTo find the kinetic energy K in terms of known quantities, we recall that for circular orbits, the gravitational force provides the centripetal force required to keep the bodies in orbit. The gravitational force is:\n\n\\[F_{\\text{gravity}} = G \\frac{M_1 M_2}{L^2}\\]\n\nAnd the centripetal force required for a mass moving in a circular orbit of radius r (where r is a portion of L relevant to each mass) and angular velocity $\\Omega$ is given by:\n\n\\[F_{\\text{centripetal}} = m \\Omega^2 r\\]\n\nFor a two-body system described, we use the reduced mass $\\mu = \\frac{M_1 M_2}{M_1 + M_2}$, which effectively experiences the centripetal acceleration due to the gravitational attraction between the two masses.\n\nThe kinetic energy, knowing it derives from the dynamics of circular motion where the gravitational force acts as the centripetal force, is given by:\n\n\\[K = \\frac{1}{2} \\mu v^2 = \\frac{1}{2} \\mu (\\Omega L)^2\\]\n\nHowever, to align kinetic and potential energy in terms of $\\mu$, G, M, and L, and explicitly realize the angular dynamics within circular motion, \n\n\\[F_{\\text{gravity}} = F_{\\text{centripetal}}\\]\n\n\\[G \\frac{M_1 M_2}{L^2} = \\mu \\Omega^2 L\\]\n\nSolving for $\\Omega^2$ and substituting back gives insight into K's dependence on these quantities, but given the task, we integrate existing understanding: for circular orbits, the kinetic energy (K) is half the magnitude but opposite in sign to the potential energy (U) in a stable gravitational two-body system, leading to:\n\n\\[K = \\frac{1}{2} G \\frac{M \\mu}{L}\\]\n\nThis relation arises from equating the gravitational force that provides the necessary centripetal force for circular motion, revealing that the kinetic energy in such a system is precisely half the potential energy's absolute value but positive, given the negative sign conventionally assigned to gravitational potential energy.\n\nSubstituting K and U into the total energy,\n\n\\[E = \\frac{1}{2} G \\frac{M \\mu}{L} - G \\frac{M \\mu}{L}\\]\n\n\\[E = -\\frac{1}{2} G \\frac{M \\mu}{L}\\]\n\nComparable with the form $E = \\beta G \\frac{M \\mu}{L}$, this implies that **$\\beta = -\\frac{1}{2}$**.

Output 1:

Assistant: 10

Figure 9: Prompt Example of Providing Complete Problem-Solving Steps That Need to Be Scored and Standard Answers. (Calculate Question)

Provide half of the problem-solving steps that need to be scored and standard answers. (GPT-4 gives 1-9)

Input 1:

System:
You are a professional physicist and you will grade answers provided by physics students by reference to standard answers. The full score is 10 points, and the minimum score is 0 points. If the student gives the final answer, full marks will be awarded directly. If the student does not give the final answer or the final answer is incorrect, please score based on the proportion of correct calculation steps given by the student. You only need to output a score number.
User:
Standard answer:
Solution: The thin layer reflects the monochromatic light of the wavelength $\\lambda$ in the best way, if the following equation holds true $ 2 n d \\cos \\beta=(2 k+1) \\frac{\\lambda}{2}, \\quad k=0,1,2, \\ldots, $ where k denotes an integer and $\\beta$ is the angle of refraction satisfying $ \\frac{\\sin \\alpha}{\\sin \\beta}=n \\text {. } $ Hence, $ \\cos \\beta=\\sqrt{1-\\sin ^2 \\beta}=\\frac{1}{n} \\sqrt{n^2-\\sin ^2 \\alpha} . $ Substituting to $2 n d \\cos \\beta=(2 k+1) \\frac{\\lambda}{2}, \\quad k=0,1,2, \\ldots$ we obtain $ 2 d \\sqrt{n^2-\\sin ^2 \\alpha}=(2 k+1) \\frac{\\lambda}{2} . $ If the white light falls on a layer, the colors of wavelengths obeying $2 d \\sqrt{n^2-\\sin ^2 \\alpha}=(2 k+1) \\frac{\\lambda}{2}$ are reinforced in the reflected light. If the wavelength of the reflected light is $\\lambda_0$, the thickness of the layer satisfies for the k th order interference $ d_k=\\frac{(2 k+1) \\lambda_0}{4 \\sqrt{n^2-\\sin ^2 \\alpha}}=(2 k+1) d_0 $ For given values and $k=0$ we obtain $d_0=1.01 \\cdot 10^{-7} \\mathrm{~m}$. a) The mass of the soap film is $m_k=\\varrho_k b h d_k$. Substituting the given values, we get $m_0=6.06 \\cdot 10^{-2} \\mathrm{mg}, m_1=18.2 \\cdot 10^{-2} \\mathrm{mg}, m_2=30.3 \\cdot 10^{-8} \\mathrm{mg}$, etc. The mass of the thinnest film thus cannot be determined by given laboratory scales.', ' If the light falls at the angle of $30^{\\circ}$ then the film seen from the perpendicular direction cannot be colored. It would appear dark.
Student answer:
To determine whether it's possible to measure the mass of the soap film using the laboratory scales with a calibration accuracy of \\(0.1 \\, \\mathrm{mg}\\), we first need to estimate the mass of the soap film. The mass of the soap film can be found if we can estimate its volume and use the given density.\n\nHowever, to check the feasibility of the calculation based on the given conditions, we need first to understand the relevance of the reflected green light (wavelength \\(\\lambda_0 = 500 \\, \\mathrm{nm}\\)) and its relation to the thickness of the soap film, because this will indirectly help us in estimating the volume.\n\n### Thickness of the Soap Film\n\nInterference of light in thin films (like soap films) can create constructive or destructive interference patterns, which depend on the thickness of the film (\\(d\\)), the wavelength of the light in the film (\\(\\lambda\\)) which is different from the wavelength in vacuum (\\(\\lambda_0\\)) due to the refractive index, and the angle of incidence (\\(\\alpha\\)).\n\nFor constructive interference of reflected light (which we assume here for seeing a strong green color), the path difference must be an integral multiple of the wavelength (\\(m\\lambda\\)), where \\(m\\) is an integer (typically starting with \\(m=1\\) for the first bright fringe). However, considering that the light needs to travel down and back through the film, for the reflected light to have a strong intensity at a particular wavelength, the condition involving the thickness of the film \\(d\\), the angle of refraction (\\(\\beta\\)), and the refractive index \\(n\\) must be considered:\n\n\\[\\delta = 2d \\cdot n \\cdot \\cos(\\beta) = m\\lambda, \\quad m=1,2,3,...\\]\n\nGiven that \\(\\lambda_0\\) is the wavelength in air and \\(n\\) is the relative refractive index, the wavelength of light in the film \\(\\lambda\\) is given by \\(\\lambda = \\lambda_0 / n\\).\n\nSnell's law relates the angles of incidence and refraction:\n\n\\[n = \\frac{\\sin(\\alpha)}{\\sin(\\beta)}\\]\n\nFor small angles or assuming the angle of refraction is relatively small, \\(\\cos(\\beta) \\approx 1\\). However, in general, we should calculate \\(\\beta\\) to be more precise.\n\nLet's simplify the calculation by focusing on the first order (\\(m=1\\)):\n\n\\[2d = \\frac{\\lambda}{n}\\]\n\nConsidering that we neglect the angle effect for simplification (\\(\\cos(\\beta) \\approx 1\\)), we can directly solve for \\(d\\), given \\(\\lambda_0\\) and \\(n\\).\n\n\\[d = \\frac{\\lambda_0}{2n} = \\frac{500 \\times 10^{-9} \\, \\mathrm{m}}{2 \\times 1.33}\\]\n\n\\[d \\approx 1.88 \\times 10^{-7} \\, \\mathrm{m}\\]\n\n### Volume and Mass of the Soap Film\n\nGiven the dimensions of the rectangle (\\(b\\) and \\(h\\)), and assuming the film has two sides, the volume (\\(V\\)) of the soap film is:\n\n\\[V = 2 \\times d \\times b \\times h\\]\n\n\\[V \\approx 2 \\times 1.88 \\times 10^{-7} \\, \\mathrm{m} \\times 0.020 \\, \\mathrm{m} \\times 0.030 \\, \\mathrm{m}\\]\n\n\\[V \\approx 2.26 \\times 10^{-9} \\, \\mathrm{m}^3\\]\n\nThe mass (\\(m\\)) of the soap film is then given by:\n\n\\[m = \\varrho \\times V\\]\n\n\\[m \\approx 1000 \\, \\mathrm{kg/m}^3 \\times 2.26 \\times 10^{-9} \\, \\mathrm{m}^3\\]\n\n\\[m \\approx 2.26 \\times 10^{-6} \\, \\mathrm{kg}\\]\n\n\\[m \\approx 2.26 \\, \\mathrm{mg}\\]\n\n### Conclusion\n\nWith a mass of approximately \\(2.26 \\, \\mathrm{mg}\\) for the soap film, and given that the scales have a calibration accuracy of \\(0.1 \\, \\mathrm{mg}\\), it is indeed possible to determine the mass of the soap film using the laboratory scales.

Output 1:

Assistant: 4

Figure 10: Prompt Example of Providing Half of the Problem-Solving Steps That Need to Be Scored and Standard Answers.

Give completely wrong problem-solving steps that need to be scored and standard answers. (GPT-4 gives 0) [1]

Input 1:

System:
You are a professional physicist and you will grade answers provided by physics students by reference to standard answers. The full score is 10 points, and the minimum score is 0 points. If the student gives the final answer, full marks will be awarded directly. If the student does not give the final answer or the final answer is incorrect, please score based on the proportion of correct calculation steps given by the student. You only need to output a score number.
User:
Standard answer:
Solution: The thin layer reflects the monochromatic light of the wavelength $\\lambda$ in the best way, if the following equation holds true $ 2 n d \\cos \\beta=(2 k+1) \\frac{\\lambda}{2}, \\quad k=0,1,2, \\ldots, $ where k denotes an integer and $\\beta$ is the angle of refraction satisfying $ \\frac{\\sin \\alpha}{\\sin \\beta}=n \\text {. } $ Hence, $ \\cos \\beta=\\sqrt{1-\\sin ^2 \\beta}=\\frac{1}{n} \\sqrt{n^2-\\sin ^2 \\alpha} . $ Substituting to $2 n d \\cos \\beta=(2 k+1) \\frac{\\lambda}{2}, \\quad k=0,1,2, \\ldots$ we obtain $ 2 d \\sqrt{n^2-\\sin ^2 \\alpha}=(2 k+1) \\frac{\\lambda}{2} . $ If the white light falls on a layer, the colors of wavelengths obeying $2 d \\sqrt{n^2-\\sin ^2 \\alpha}=(2 k+1) \\frac{\\lambda}{2}$ are reinforced in the reflected light. If the wavelength of the reflected light is $\\lambda_0$, the thickness of the layer satisfies for the k th order interference $ d_k=\\frac{(2 k+1) \\lambda_0}{4 \\sqrt{n^2-\\sin ^2 \\alpha}}=(2 k+1) d_0 $ For given values and $k=0$ we obtain $d_0=1.01 \\cdot 10^{-7} \\mathrm{~m}$. a) The mass of the soap film is $m_k=\\varrho_k b h d_k$. Substituting the given values, we get $m_0=6.06 \\cdot 10^{-2} \\mathrm{mg}, m_1=18.2 \\cdot 10^{-2} \\mathrm{mg}, m_2=30.3 \\cdot 10^{-8} \\mathrm{mg}$, etc. The mass of the thinnest film thus cannot be determined by given laboratory scales.', ' If the light falls at the angle of $30^{\\circ}$ then the film seen from the perpendicular direction cannot be colored. It would appear dark.
Student answer:
Hello! World.

Output 1:

Assistant: 0

Figure 11: Prompt example of Providing Completely Wrong Problem-Solving Steps that Need to Be Scored and Standard Answers. (1)

Give completely wrong problem-solving steps that need to be scored and standard answers. (GPT-4 gives 0) [2]

Input 1:

System:
You are a professional physicist and you will grade answers provided by physics students by reference to standard answers. The full score is 10 points, and the minimum score is 0 points. If the student gives the final answer, full marks will be awarded directly. If the student does not give the final answer or the final answer is incorrect, please score based on the proportion of correct calculation steps given by the student. You only need to output a score number.
User:
Standard answer:
Solution: The thin layer reflects the monochromatic light of the wavelength $\\lambda$ in the best way, if the following equation holds true $ 2 n d \\cos \\beta=(2 k+1) \\frac{\\lambda}{2}, \\quad k=0,1,2, \\ldots, $ where k denotes an integer and $\\beta$ is the angle of refraction satisfying $ \\frac{\\sin \\alpha}{\\sin \\beta}=n \\text {. } $ Hence, $ \\cos \\beta=\\sqrt{1-\\sin ^2 \\beta}=\\frac{1}{n} \\sqrt{n^2-\\sin ^2 \\alpha} . $ Substituting to $2 n d \\cos \\beta=(2 k+1) \\frac{\\lambda}{2}, \\quad k=0,1,2, \\ldots$ we obtain $ 2 d \\sqrt{n^2-\\sin ^2 \\alpha}=(2 k+1) \\frac{\\lambda}{2} . $ If the white light falls on a layer, the colors of wavelengths obeying $2 d \\sqrt{n^2-\\sin ^2 \\alpha}=(2 k+1) \\frac{\\lambda}{2}$ are reinforced in the reflected light. If the wavelength of the reflected light is $\\lambda_0$, the thickness of the layer satisfies for the k th order interference $ d_k=\\frac{(2 k+1) \\lambda_0}{4 \\sqrt{n^2-\\sin ^2 \\alpha}}=(2 k+1) d_0 $ For given values and $k=0$ we obtain $d_0=1.01 \\cdot 10^{-7} \\mathrm{~m}$. a) The mass of the soap film is $m_k=\\varrho_k b h d_k$. Substituting the given values, we get $m_0=6.06 \\cdot 10^{-2} \\mathrm{mg}, m_1=18.2 \\cdot 10^{-2} \\mathrm{mg}, m_2=30.3 \\cdot 10^{-8} \\mathrm{mg}$, etc. The mass of the thinnest film thus cannot be determined by given laboratory scales.', ' If the light falls at the angle of $30^{\\circ}$ then the film seen from the perpendicular direction cannot be colored. It would appear dark.
Student answer:
The color seen in a soap film is due to constructive and destructive interference patterns of light reflecting off the front and back surfaces of the film. The thinnest possible soap film that still displays a color will be when the film creates the condition for constructive interference for the shortest visible wavelength of light (violet) in the first order of interference (where \\(m=1\\)), observed perpendicularly (\\(\\alpha = 0\\), meaning \\(\\beta = 0\\) and thus \\(\\cos(\\beta) = 1\\)).\n\n### Constructive Interference Condition\n\nFor constructive interference from a thin film observed perpendicularly (\\(\\alpha=0^{\\circ}\\) and hence \\(\\beta=0^{\\circ}\\)), the condition for the \\(m\\)-th order of maxima is:\n\n\\[\\delta = 2d \\cdot n = m\\lambda, \\quad m=1,2,3...\\]\n\nWhere:\n- \\(\\delta\\) is the path difference that causes constructive interference.\n- \\(d\\) is the thickness of the film.\n- \\(n\\) is the refractive index of the film.\n- \\(\\lambda\\) is the wavelength of light in the film.\n- \\(m\\) is the order of interference.\n\nFor the thinnest possible film displaying color, \\(m=1\\) (first order), and we aim to solve for \\(d\\) (thickness).\n\n### Wavelength of Light in the Film\n\nThe wavelength of light in the film (\\(\\lambda\\)) is related to the wavelength in the vacuum (\\(\\lambda_0\\)) by:\n\n\\[\\lambda = \\frac{\\lambda_0}{n}\\]\n\nSince we are interested in visible light and the thinnest film would theoretically reflect violet light (since violet has the shortest wavelength in the visible spectrum, around \\(400 \\, \\mathrm{nm}\\)), we would use \\(\\lambda_0 \\approx 400 \\, \\mathrm{nm}\\).\n\n### Solving for the Minimum Thickness \\(d\\)\n\nSet \\(m=1\\) for the first order of maximum (bright fringe) and solve for \\(d\\) for violet light to find the thinnest film:\n\n\\[2d \\cdot n = \\lambda\\]\n\n\\[2d = \\frac{\\lambda_0}{n}\\]\n\nGiven that \\(\\lambda_0 \\approx 400 \\, \\mathrm{nm}\\) for violet light and substituting, we get:\n\n\\[d = \\frac{400 \\times 10^{-9} \\, \\mathrm{m}}{2n}\\]\n\nBecause the refractive index \\(n = 1.33\\) (for the soap film, similar to water), we substitute \\(n\\) into the equation:\n\n\\[d = \\frac{400 \\times 10^{-9} \\, \\mathrm{m}}{2 \\times 1.33}\\]\n\n\\[d \\approx 150.38 \\times 10^{-9} \\, \\mathrm{m}\\]\n\n\\[d \\approx 150 \\, \\mathrm{nm}\\]\n\n### Conclusion\n\nThe thinnest possible soap film that still displays a color when viewed perpendicularly would reflect violet light, characteristic of the first order of constructive interference, with a minimum thickness of approximately \\(150 \\, \\mathrm{nm}\\). This result follows from the constructive interference condition, considering the smallest possible constructive interference for visible light occurs with violet light due to its shortest wavelength in the visible spectrum.

Output 1:

Assistant: 0

Figure 12: Prompt example of Providing Completely Wrong Problem-Solving Steps that Need to Be Scored and Standard Answers. (2)

Slightly change the final answer but the error is within the reasonable range required by the question. (GPT-4 gives 10)

Input 1:

System:
You are a professional physicist and you will grade answers provided by physics students by reference to standard answers. The full score is 10 points, and the minimum score is 0 points. If the student gives the final answer, full marks will be awarded directly. If the student does not give the final answer or the final answer is incorrect, please score based on the proportion of correct calculation steps given by the student. You only need to output a score number.
User:
Standard answer:
$v = 20 \\pm 0.8 \\, \\text{m/s}$
Student answer:
20.6

Output 1:

Assistant: 10

Figure 13: Prompt Example of Slightly Change the Final Answer but the Error Is Within the Reasonable Range Required by the Question.

Give answers in normal format, Latex format, and answers without units but the numbers are the same and standard answers. (GPT-4 gives 10)

Input 1:

System:
You are a professional physicist and you will grade answers provided by physics students by reference to standard answers. The full score is 10 points, and the minimum score is 0 points. If the student gives the final answer, full marks will be awarded directly. If the student does not give the final answer or the final answer is incorrect, please score based on the proportion of correct calculation steps given by the student. You only need to output a score number.
User:
Standard answer:
$3 \\text{cm}$
Student answer:
3 \\text{cm}

Output 1:

Assistant: 10

Input 2:

System:
You are a professional physicist and you will grade answers provided by physics students by reference to standard answers. The full score is 10 points, and the minimum score is 0 points. If the student gives the final answer, full marks will be awarded directly. If the student does not give the final answer or the final answer is incorrect, please score based on the proportion of correct calculation steps given by the student. You only need to output a score number.
User:
Standard answer:
$3 \\text{cm}$
Student answer:
3cm

Output 2:

Assistant: 10

Input 3:

System:
You are a professional physicist and you will grade answers provided by physics students by reference to standard answers. The full score is 10 points, and the minimum score is 0 points. If the student gives the final answer, full marks will be awarded directly. If the student does not give the final answer or the final answer is incorrect, please score based on the proportion of correct calculation steps given by the student. You only need to output a score number.
User:
Standard answer:
$3 \\text{cm}$
Student answer:
3

Output 3:

Assistant: 10

Figure 14: Prompt Example of Providing Answers in Normal Format, Latex Format, and Answers Without Units but the Numbers Are the Same and Standard Answers.

Provide complete problem-solving steps that need to be scored with changing the final answer to a totally wrong one and standard answers. (GPT-4 gives 9)

Input 1:

System:
You are a professional physicist and you will grade answers provided by physics students by reference to standard answers. The full score is 10 points, and the minimum score is 0 points. If the student gives the final answer, full marks will be awarded directly. If the student does not give the final answer or the final answer is incorrect, please score based on the proportion of correct calculation steps given by the student. You only need to output a score number.
User:
Standard answer:
Energy (3) of the question sheet can be interpreted as describing a system of a mass $\\mu$ in a circular orbit with angular velocity $\\Omega$, radius L, around a mass M (at rest). Equating the gravitational acceleration to the centripetal acceleration:$G \\frac{M}{L^2}=\\Omega^2 L$. This is indeed Kepler's third law (for circular orbits). Then, from (7), $E=-\\frac{1}{2} G \\frac{M \\mu}{L}$.**$\\beta=-\\frac{1}{2}$**.
Student answer:
To simplify the total energy expression and arrive at a formula such as $E = \\beta G \\frac{M \\mu}{L}$, we need to incorporate both kinetic and potential energy terms and their relation in a two-body system of circular orbits. \n\nFrom previous considerations, the total energy E of the system combines both kinetic (K) and potential (U) energies:\n\n\\[E = K + U\\]\n\nFrom the provided equation and context, the potential energy of two bodies attracting each other gravitationally is:\n\n\\[U = -G \\frac{M \\mu}{L}\\]\n\nTo find the kinetic energy K in terms of known quantities, we recall that for circular orbits, the gravitational force provides the centripetal force required to keep the bodies in orbit. The gravitational force is:\n\n\\[F_{\\text{gravity}} = G \\frac{M_1 M_2}{L^2}\\]\n\nAnd the centripetal force required for a mass moving in a circular orbit of radius r (where r is a portion of L relevant to each mass) and angular velocity $\\Omega$ is given by:\n\n\\[F_{\\text{centripetal}} = m \\Omega^2 r\\]\n\nFor a two-body system described, we use the reduced mass $\\mu = \\frac{M_1 M_2}{M_1 + M_2}$, which effectively experiences the centripetal acceleration due to the gravitational attraction between the two masses.\n\nThe kinetic energy, knowing it derives from the dynamics of circular motion where the gravitational force acts as the centripetal force, is given by:\n\n\\[K = \\frac{1}{2} \\mu v^2 = \\frac{1}{2} \\mu (\\Omega L)^2\\]\n\nHowever, to align kinetic and potential energy in terms of $\\mu$, G, M, and L, and explicitly realize the angular dynamics within circular motion, \n\n\\[F_{\\text{gravity}} = F_{\\text{centripetal}}\\]\n\n\\[G \\frac{M_1 M_2}{L^2} = \\mu \\Omega^2 L\\]\n\nSolving for $\\Omega^2$ and substituting back gives insight into K's dependence on these quantities, but given the task, we integrate existing understanding: for circular orbits, the kinetic energy (K) is half the magnitude but opposite in sign to the potential energy (U) in a stable gravitational two-body system, leading to:\n\n\\[K = \\frac{1}{2} G \\frac{M \\mu}{L}\\]\n\nThis relation arises from equating the gravitational force that provides the necessary centripetal force for circular motion, revealing that the kinetic energy in such a system is precisely half the potential energy's absolute value but positive, given the negative sign conventionally assigned to gravitational potential energy.\n\nSubstituting K and U into the total energy,\n\n\\[E = \\frac{1}{2} G \\frac{M \\mu}{L} - G \\frac{M \\mu}{L}\\]\n\n\\[E = -\\frac{1}{2} G \\frac{M \\mu}{L}\\]\n\nComparable with the form $E = \\beta G \\frac{M \\mu}{L}$, this implies that **$\\beta = -\\frac{1}{99}$**.

Output 1:

Assistant: 9

Figure 15: Prompt example of Providing Complete Problem-Solving Steps that Need to Be Scored with Changing the Final Answer to a Totally Wrong One and Standard Answers.

Cut a candidate's zero marks solution to a half but with the correct final answer. (GPT-4 gives 0)

Input 1:

System:
You are a professional physicist and you will grade answers provided by physics students by reference to standard answers. The full score is 10 points, and the minimum score is 0 points. If the student gives the final answer, full marks will be awarded directly. If the student does not give the final answer or the final answer is incorrect, please score based on the proportion of correct calculation steps given by the student. You only need to output a score number.
User:
Standard answer:
Energy (3) of the question sheet can be interpreted as describing a system of a mass $\\mu$ in a circular orbit with angular velocity $\\Omega$, radius L, around a mass M (at rest). Equating the gravitational acceleration to the centripetal acceleration:$G \\frac{M}{L^2}=\\Omega^2 L$. This is indeed Kepler's third law (for circular orbits). Then, from (7), $E=-\\frac{1}{2} G \\frac{M \\mu}{L}$.**$\\beta=-\\frac{1}{2}$.**
Student answer:
To determine whether it's possible to measure the mass of the soap film using the laboratory scales with a calibration accuracy of \\(0.1 \\, \\mathrm{mg}\\), we first need to estimate the mass of the soap film. The mass of the soap film can be found if we can estimate its volume and use the given density.\n\nHowever, to check the feasibility of the calculation based on the given conditions, we need first to understand the relevance of the reflected green light (wavelength \\(\\lambda_0 = 500 \\, \\mathrm{nm}\\)) and its relation to the thickness of the soap film, because this will indirectly help us in estimating the volume.\n\n### Thickness of the Soap Film\n\nInterference of light in thin films (like soap films) can create constructive or destructive interference patterns, which depend on the thickness of the film (\\(d\\)), the wavelength of the light in the film (\\(\\lambda\\)) which is different from the wavelength in vacuum **$\\beta = -\\frac{1}{2}$.**

Output 1:

Assistant: 9

Figure 16: Prompt example of Cutting a Candidate's Zero Marks Solution to a Half but With the Correct Final Answer.

Cut a candidate's full marks solution to a half and add the correct final answer. (GPT-4 gives 10; Should 5-7)

Input 1:

System:
You are a professional physicist and you will grade answers provided by physics students by reference to standard answers. The full score is 10 points, and the minimum score is 0 points. If the student gives the final answer, full marks will be awarded directly. If the student does not give the final answer or the final answer is incorrect, please score based on the proportion of correct calculation steps given by the student. You only need to output a score number.
User:
Standard answer:
Energy (3) of the question sheet can be interpreted as describing a system of a mass $\\mu$ in a circular orbit with angular velocity $\\Omega$, radius L, around a mass M (at rest). Equating the gravitational acceleration to the centripetal acceleration:$G \\frac{M}{L^2}=\\Omega^2 L$. This is indeed Kepler's third law (for circular orbits). Then, from (7), $E=-\\frac{1}{2} G \\frac{M \\mu}{L}$.**$\\beta=-\\frac{1}{2}$**.
Student answer:
To simplify the total energy expression and arrive at a formula such as $E = \\beta G \\frac{M \\mu}{L}$, we need to incorporate both kinetic and potential energy terms and their relation in a two-body system of circular orbits. \n\nFrom previous considerations, the total energy E of the system combines both kinetic (K) and potential (U) energies:\n\n\\[E = K + U\\]\n\nFrom the provided equation and context, the potential energy of two bodies attracting each other gravitationally is:\n\n\\[U = -G \\frac{M \\mu}{L}\\]\n\nTo find the kinetic energy K in terms of known quantities, we recall that for circular orbits, the gravitational force provides the centripetal force required to keep the bodies in orbit. The gravitational force is:\n\n\\[F_{\\text{gravity}} = G \\frac{M_1 M_2}{L^2}\\]\n\nAnd the centripetal force required for a mass moving in a circular orbit of radius r (where r is a portion of L relevant to each mass) and angular velocity $\\Omega$ is given by:\n\n\\[F_{\\text{centripetal}} = m \\Omega^2 r\\]\n\nFor a two-body system described, we use the reduced mass $\\mu = \\frac{M_1 M_2}{M_1 + M_2}$, which effectively experiences the centripetal acceleration due to the gravitational attraction between the two masses.\n\nThe kinetic energy, knowing it derives from the dynamics of circular motion where the gravitational force acts as the centripetal force, **$\\beta = -\\frac{1}{2}$.**

Output 1:

Assistant: 10

Figure 17: Prompt example of Cutting a Candidate's Full Marks Solution to a Half and Add the Correct Final Answer.

Randomly delete some scoring points from a candidate's full marks solution, but the final answer is correct (GPT-4 gives 10; Should 1-9)

Input 1:

System:
You are a professional physicist and you will grade answers provided by physics students by reference to standard answers. The full score is 10 points, and the minimum score is 0 points. If the student gives the final answer, full marks will be awarded directly. If the student does not give the final answer or the final answer is incorrect, please score based on the proportion of correct calculation steps given by the student. You only need to output a score number.
User:
Standard answer:
Energy (3) of the question sheet can be interpreted as describing a system of a mass $\\mu$ in a circular orbit with angular velocity $\\Omega$, radius L, around a mass M (at rest). Equating the gravitational acceleration to the centripetal acceleration:$G \\frac{M}{L^2}=\\Omega^2 L$. This is indeed Kepler's third law (for circular orbits). Then, from (7), $E=-\\frac{1}{2} G \\frac{M \\mu}{L}$.**$\\beta=-\\frac{1}{2}$**.
Student answer:
To simplify the total energy expression and arrive at a formula such as $E = \\beta G \\frac{M \\mu}{L}$, we need to incorporate both kinetic and potential energy terms and their relation in a two-body system of circular orbits. \n\nFrom previous considerations, the total energy E of the system combines both kinetic (K) and potential (U) energies:
we use the reduced mass $\\mu = \\frac{M_1 M_2}{M_1 + M_2}$, which effectively experiences the centripetal acceleration due to the gravitational attraction between the two masses.\n\nThe kinetic energy, knowing it derives from the dynamics of circular motion where the gravitational force acts as the centripetal force, is given by:\n\n\\[K = \\frac{1}{2} \\mu v^2 = \\frac{1}{2} \\mu (\\Omega L)^2\\]
the kinetic energy (K) is half the magnitude but opposite in sign to the potential energy (U) in a stable gravitational two-body system, leading to:\n\n\\[K = \\frac{1}{2} G \\frac{M \\mu}{L}\\]\n\nThis relation arises from equating the gravitational force that provides the necessary centripetal force for circular motion,
$\\beta = -\\frac{1}{99}$.

Output 1:

Assistant: 10

Figure 18: Prompt Example of Randomly Deleting Some Scoring Points from a Candidate's Full Marks Solution, but the Final Answer Is Correct.

A.3 Modification of Identifiers

We have identify and amend a series of identifiers, which include but are not limited to the following:

1	2	3	4	5	6	7	8	9
a	b	c	d	e	f	g	h	i
a)	b)	c)	d)	e)	f)	g)	h)	i)
A.	B.	C.	D.	E.	F.	G.	H.	I.
A.1	A.2	A.3	A.4	A.5	A.6	A.7	A.8	A.9
I	II	III	IV	V	VI	VII	VIII	IX
i	ii	iii	iv	v	vi	vii	viii	ix

Table 3: Modified identifiers

A.4 Prompt Examples

In this section, we display the prompt examples of how we prompt LLMs to inference physical questions, mark solutions.

Prompt example for LLMs inference

Input 1:

System: You are a professional physicist, and you will answer the following physics questions. The mathematical formulas involved in the questions are expressed in Latex code.
User: During the Soviet-French experiment on the optical location of the Moon the light pulse of a ruby laser $(\\lambda=0,69 \\mu \\mathrm{m})$ was directed to the Moon's surface by the telescope with a diameter of
......
$d_p=5 \\mathrm{~mm}$. Plank constant is $\\mathrm{h}=6.610^{-34} \\mathrm{Js}$. What must the accuracy to direct the telescope optical axis be in this experiment?

Output 1:

Assistant: To determine the required accuracy to direct the telescope optical axis in this experiment, we need to consider the size of the spot that the laser beam forms on the Moon's surface.\n\nThe size of the spot can be estimated using the formula for the diffraction-limited spot size, given by:\n\n\\begin{equation}\n\\theta \\approx \\frac{1.22 \\lambda}{D}\n\\end{equation}\n\nwhere $\\theta$ is the angular size of the spot, $\\lambda$
......
is the size of the spot on the photodetector and d_p is the diameter of the pupil of the eye.\n\n\\begin{equation}\n\\Delta x \\approx (3.25 \\times 10^{-7}) \\times (5 \\times 10^{-3})\n\\end{equation}\n\n\\begin{equation}\n\\Delta x \\approx 1.625 \\times 10^{-9} \\mathrm{~m}\n\\end{equation}\n\nTherefore, the required accuracy to direct the telescope optical axis in this experiment is approximately $1.625 \\times 10^{-9} \\mathrm{~m}$.

Input 2:

System: You are a professional physicist, and you will answer the following physics questions. The mathematical formulas involved in the questions are expressed in Latex code.
User: During the Soviet-French experiment on the optical location of the Moon the light pulse of a ruby laser $(\\lambda=0,69 \\mu \\mathrm{m})$ was directed to the Moon's surface by the telescope with a diameter of
......
$d_p=5 \\mathrm{~mm}$. Plank constant is $\\mathrm{h}=6.610^{-34} \\mathrm{Js}$. What must the accuracy to direct the telescope optical axis be in this experiment?
Assistant: To determine the required accuracy to direct the telescope optical axis in this experiment, we need to consider the size of the spot that the laser beam forms on the Moon's surface.\n\nThe size of the spot can be estimated using the formula for the diffraction-limited spot size, given by:\n\n\\begin{equation}\n\\theta \\approx \\frac{1.22 \\lambda}{D}\n\\end{equation}\n\nwhere $\\theta$ is the angular size of the spot, $\\lambda$
......
is the size of the spot on the photodetector and d_p is the diameter of the pupil of the eye.\n\n\\begin{equation}\n\\Delta x \\approx (3.25 \\times 10^{-7}) \\times (5 \\times 10^{-3})\n\\end{equation}\n\n\\begin{equation}\n\\Delta x \\approx 1.625 \\times 10^{-9} \\mathrm{~m}\n\\end{equation}\n\nTherefore, the required accuracy to direct the telescope optical axis in this experiment is approximately $1.625 \\times 10^{-9} \\mathrm{~m}$.

Output 2:

Assistant: To determine the part of the emitted laser energy that can be detected after reflection on the Moon, we need to consider the reflectance of the mirror on the Moon's surface and the losses due to the telescope's collecting
......
\\times 10^8 / 0.69 \\times 10^{-6})} \\mathrm{~quanta}\n\\end{equation}\n\n\\begin{equation}\nn \\approx 7.246 \\times 10^{25} \\mathrm{~quanta}\n\\end{equation}\n\nGiven that the threshold sensitivity of the eye is $n = 100$ light quanta, we can see that the energy of the laser pulse is much larger than the threshold sensitivity. Therefore, we can expect that the reflected light pulse would be visible with the naked eye.

Input 3:

......

Figure 19: Prompt example of pure text inference

Marking Examples

Input 1:

System: You are a professional physicist and you will grade answers provided by physics students by reference to standard answers. The full score is 10 points, and the minimum score is 0 points. If the student gives the final answer, full marks will be awarded directly. If the student does not give the final answer or the final answer is incorrect, please score based on the proportion of correct calculation steps given by the student. You only need to output a score number.
User: Standard answer: 10 m/s Student answer: 10

Output 1:

Assistant: 10

Input 2:

System: You are a professional physicist and you will grade answers provided by physics students by reference to standard answers. The full score is 10 points, and the minimum score is 0 points. If the student gives the final answer, full marks will be awarded directly. If the student does not give the final answer or the final answer is incorrect, please score based on the proportion of correct calculation steps given by the student. You only need to output a score number.
User: Standard answer: Start with $\\Delta x=\\frac{1}{2} a t^2+v_i t$. We have four times. $t_0=0$ is when the train starts, and when the first car is aligned with the observer. t_1 is when the end of the first car is aligned with the observer. Then $L=\\frac{1}{2} a t_1{ }^2$. We are assuming the car has a length $L. t_2$ is when the tenth car is first aligned with the observer, so $9 L=\\frac{1}{2} a t_2{ }^2$ and finally, t_3 is when that car has passed, $10 L=\\frac{1}{2} a t_3{ }^2$. From the equation for t_1 we find $2 L / a=25 \\mathrm{~s}^2$ so $t_2=\\sqrt{9 \\cdot 25 \\mathrm{~s}^2}=15 \\mathrm{~s}$ and $t_3=\\sqrt{10 \\cdot 25 \\mathrm{~s}^2}=15.81 \\mathrm{~s}$. Student answer: The speed of the center of mass (CM) of a system can be found using the formula for the velocity of the center of mass \\(v_{CM}\\), which is given by:\n\n\\[\nv_{CM} = \\frac{m_1v_1 + m_2v_2}{m_1 + m_2}\n\\]\n\nWhere:\n- \\(m_1, m_2\\) are the masses of the two objects, and\n- \\(v_1, v_2\\) are their velocities.\n\nGiven:\n- \\(m_1 = 3 \\, \\text{kg}\\) and \\(v_1 = 30 \\, \\text{m/s}\\) (to the right, which we can consider as positive direction),\n- \\(m_2 = 2 \\, \\text{kg}\\) and \\(v_2 = -20 \\, \\text{m/s}\\) (to the left, which we consider as negative direction because it is opposite to \\(v_1\\)),\n\nSubstituting the given values into the formula for \\(v_{CM}\\):\n\n\\[\nv_{CM} = \\frac{(3 \\, \\text{kg})(30 \\, \\text{m/s}) + (2 \\, \\text{kg})(-20 \\, \\text{m/s})}{3 \\, \\text{kg} + 2 \\, \\text{kg}}\n\\]\n\n\\[\nv_{CM} = \\frac{90 - 40}{5}\n\\]\n\n\\[\nv_{CM} = \\frac{50}{5}\n\\]\n\n\\[\nv_{CM} = 10 \\, \\text{m/s}\n\\]\n\nTherefore, after the collision, the center of mass of the system is moving at a speed of \\(10 \\, \\text{m/s}\\) to the right.

Output 2:

Assistant: 0

Input 3:

System: You are a professional physicist and you will grade answers provided by physics students by reference to standard answers. The full score is 10 points, and the minimum score is 0 points. If the student gives the final answer, full marks will be awarded directly. If the student does not give the final answer or the final answer is incorrect, please score based on the proportion of correct calculation steps given by the student. You only need to output a score number.
User: Standard answer: The thin layer reflects the monochromatic light of the wavelength $\\lambda$ in the best way, if the following equation holds true $ 2 n d \\cos \\beta=(2 k+1) \\frac{\\lambda}{2}, \\quad k=0,1,2, \\ldots, $ where
......
\\mathrm{~m}$. a) The mass of the soap film is $m_k=\\varrho_k b h d_k$. Substituting the given values, we get $m_0=6.06 \\cdot 10^{-2} \\mathrm{mg}, m_1=18.2 \\cdot 10^{-2} \\mathrm{mg}, m_2=30.3 \\cdot 10^{-8} \\mathrm{mg}$, etc. The mass of the thinnest film thus cannot be determined by given laboratory scales.

Student answer: To determine whether it's possible to measure the mass of the soap film using the laboratory scales with a calibration accuracy of \\(0.1 \\, \\mathrm{mg}\\), we first need to estimate the mass of the soap film. The
......
approximately \\(2.26 \\, \\mathrm{mg}\\) for the soap film, and given that the scales have a calibration accuracy of \\(0.1 \\, \\mathrm{mg}\\), it is indeed possible to determine the mass of the soap film using the laboratory scales.

Output 3:

Assistant: 3

Figure 20: Marking Examples

A.5 Gemini Output Examples

Questions: In a simple model, an atomic nucleus can be thought of as a ball consisting of closely packed nucleons [see Fig. 1(a)], where the nucleons are hard balls of radius $r_N = 0.85\text{fm}\left(1\text{fm} = 10^{-15}\text{ m}\right)$. The nuclear force is present only for two nucleons in contact. The volume of the nucleus V is larger than the volume of all nucleons AV_N, where $V_N = \frac{4}{3}r_N^3\pi$. The ratio $f = AV_N/V$ is called the packing factor and gives the percentage of space filled by the nuclear matter. Calculate what would be the packing factor f if nucleons were arranged in a "simple cubic" (SC) crystal system, where each nucleon is centered on a lattice point of an infinite cubic lattice [see Fig. 1(b)]. (0.3 points)

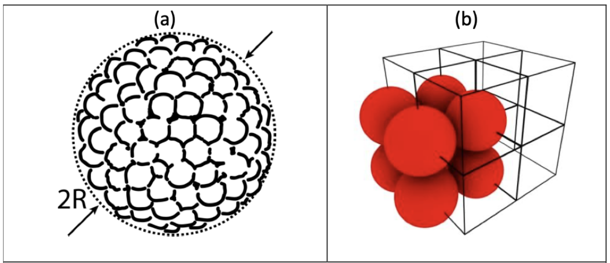


Fig. 1. (a) An atomic nucleus as a ball of closely packed nucleons. (b) The SC packing.

Standard Answer: In the SC-system, in each of 8 corners of a given cube there is one unit (atom, nucleon, etc.), but it is shared by 8 neighboring cubes - this gives a total of one nucleon per cube. If nucleons are touching, as we assume in our simplified model, then $a = 2r_N$ is the cube edge length a. The volume of one nucleon is then $V_N = \frac{4}{3}r_N^3\pi = \frac{4}{3}\left(\frac{a}{2}\right)^3\pi = \frac{4a^3}{3\cdot 8}\pi = \frac{\pi}{6}a^3$ from which we obtain $f = \frac{V_N}{a^3} = \frac{\pi}{6} \approx 0.52$

ChatGPT-4-Vision Output:(Score: 10) In the "simple cubic" (SC) crystal system, each nucleon is centered on a lattice point, and because we're assuming the nucleons to be hard spheres, the sides of the cubic lattice cell must be twice the radius of a nucleon to accommodate one nucleon inside without overlapping with others. Given that the radius of a nucleon r_N is 0.85 fm (fermi), which is 0.85×10^{-15} meters, the side length a of the cubic cell will be:$a = 2 \times r_N$ The volume V_N of a single nucleon (since it's a sphere) is:$V_N = \frac{4}{3}\pi r_N^3$ Now, the volume V_{SC} of the cubic cell which contains one nucleon would be: $V_{SC} = a^3 = (2r_N)^3 = 8r_N^3$ The packing factor f will be the ratio of the nucleon's volume to the volume of the cubic cell: $f = \frac{AV_N}{V} = \frac{V_N}{V_{SC}}$ Since each cell volume V_{SC} contains one nucleon for the simple cubic packing, we don't need to multiply V_N by A, because A will cancel when we take the ratio for one cubic cell. Plugging in the expressions for V_N and V_{SC} gives: $f = \frac{\frac{4}{3}\pi r_N^3}{8r_N^3}$ alculating this gives:$f = \frac{\pi}{6}$ Numerically, this evaluates to approximately:$f \approx \frac{3.14159}{6} \approx 0.5236$Or, in percentage terms:$f \approx 52.36\%$his means the packing factor or the percentage of space filled by the nuclear matter in a simple cubic arrangement is approximately 52.36%.

Gemini-Pro-Vision Output:(Score: 2) I see the image you sent and can use it to understand the setup of the Kelvin water dropper. The absolute value of the charge, Q_0, of each droplet as they separate from the tubes and at the instant the capacitor's charge is q can be expressed as: $Q_0 = Cq/\left(2r_{\max}\right)$ where: C is the capacitance. q is the charge on the capacitor. $r_{\max}$ is the maximum radius of the droplet as it breaks free from the nozzle This expression is derived based on the following: 1. Charge Quantization: Each droplet acquires a quantized charge, meaning its charge is a multiple of the fundamental electron charge (e). As the first droplet falls, it acquires a small charge imbalance, which is then shared equally between the two electrodes due to the capacitor. 4. Maximum Radius: The maximum radius, $r_{\max}$, relates to the amount of charge a droplet can hold due to its surface tension. A larger droplet can hold more charge. Therefore, the absolute value of the charge on each droplet depends on the capacitance, the capacitor's charge, and the maximum radius the droplet can achieve before breaking free. It's important to note that this expression neglects the effect mentioned in Part A-iii, where the charge on the droplet can induce a charge on the metal nozzle. This additional effect can slightly modify the charge acquired by the droplet. I hope this helps! Let me know if you have any other questions.

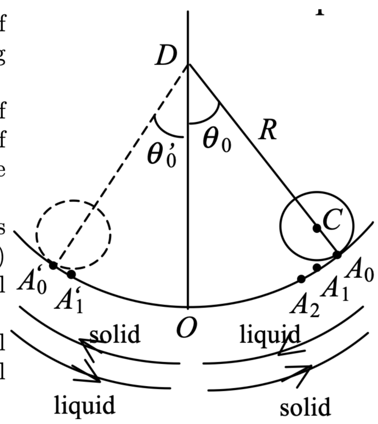
Consider a sphere filled with liquid inside rolling back and forth at the bottom of a spherical bowl. $\cdots$ please study the movement of the sphere for the following three cases:

1. **W** behaves as in ideal solid state, meanwhile **W** contacts the inner wall of the spherical shell so closely that they can be taken as solid sphere as a whole of radius r with an abrupt density change across the interface between the inside wall of the shell and **W**.

(1) Calculate the rotational inertia I of the sphere with respect to the axis passing through its center C. (You are asked to show detailed steps.) (1.0 points)

(2) Calculate the period T_1 of the sphere rolling back and forth with a small amplitude without slipping at the bottom of the spherical bowl. (2.5 points)

2. **W** behaves as an ideal liquid with no friction between **W** and the spherical shell. Calculate the period T_2 of the sphere rolling back and forth with a small amplitude without slipping at the bottom of the spherical bowl. (2.5 points)



(a) An Original Question with Second Order Questions in PhoPile

```
Question: Consider a sphere filled with liquid inside rolling back and forth at
    the bottom of a spherical bowl. ... please study the movement of the sphere
    for the following three cases: $\mathbf{W}$ behaves as in ideal solid state,
     meanwhile $\mathbf{W}$ contacts the inner wall of the spherical shell so
    closely that they can be taken as solid sphere as a whole of radius $r$ with
     an abrupt density change across the interface between the inside wall of
    the shell and $\mathbf{W}$.
Calculate the rotational inertia $I$ of the sphere with respect to the axis
    passing through its center $C$. (You are asked to show detailed steps.)
Question number: 1
Sub-question number: 1
Sub-sub question number: 1
Solution: ...
Question image path: ./pic/question/34.png
Answer image path: ....

Question: Calculate the period $T_1$ of the sphere rolling back and forth with a
     small amplitude without slipping at the bottom of the spherical bowl.
Question number: 1
Sub-question number: 1
Sub-sub question number: 2
Solution: ...
Question image path:...
Answer image path: ....

Question: $\mathbf{W}$ behaves as an ideal liquid with no friction between $\
    mathbf{W}$ and the spherical shell. Calculate the period $T_2$ of the sphere
     rolling back and forth with a small amplitude without slipping at the
    bottom of the spherical bowl.
Question number: 1
Sub-question number: 2
Sub-sub-question number: null
Solution: ...
Question image path:...
Answer image path: ....
```

(b) The Corresponding Question in PhoPile: We delete '1.', '(1)', and '(2) in text. At the same time, we input the Arabic number to 'Question number', 'Sub-question number', and 'Sub-sub-question number' key.

Figure 21: Demonstration of Question Index Processing