Benchmarking Foundation Models with Retrieval-Augmented Generation in Olympic-Level Physics Problem Solving

Shunfeng Zheng[1]*, Yudi Zhang[2]*, Meng Fang[3], Zihan Zhang[1], Zhitan Wu[4], Mykola Pechenizkiy[2], Ling Chen[1]

[1]AAII, University of Technology Sydney, New South Wales, Australia
[2]Eindhoven University of Technology, Eindhoven, The Netherlands
[3]University of Liverpool, Liverpool, United Kingdom
[4]University of New South Wales, New South Wales, Australia
Shunfeng.Zheng@student.uts.edu.au, y.zhang5@tue.nl, Meng.Fang@liverpool.ac.uk
Zihan.Zhang-5@student.uts.edu.au, zhitan.wu@student.unsw.edu.au, m.pechenizkiy@tue.nl
Ling.Chen@uts.edu.au

Abstract

Retrieval-augmented generation (RAG) with foundation models has achieved strong performance across diverse tasks, but their capacity for expert-level reasoning—such as solving Olympiad-level physics problems—remains largely unexplored. Inspired by the way students prepare for competitions by reviewing past problems, we investigate the potential of RAG to enhance physics reasoning in foundation models. We introduce **PhoPile**, a high-quality multimodal dataset specifically designed for Olympiad-level physics, enabling systematic study of retrieval-based reasoning. PhoPile includes diagrams, graphs, and equations, capturing the inherently multimodal nature of physics problem solving. Using PhoPile, we benchmark RAG-augmented foundation models, covering both large language models (LLMs) and large multimodal models (LMMs) with multiple retrievers. Our results demonstrate that integrating retrieval with physics corpora can improve model performance, while also highlighting challenges that motivate further research in retrieval-augmented physics reasoning.

1 Introduction

Physics plays a foundational role in natural sciences and engineering, underpinning progress in fields ranging from construction and aerospace to electronics and materials science (Serway et al., 2000). Mastering physics requires not only conceptual understanding of natural laws but also the ability to integrate them with quantitative analysis, diagrams, and symbolic reasoning. Recent years have witnessed the exceptional performance of foundation models, including large language models (LLMs) and large multimodal models (LMMs), such as GPT-3 (Brown et al., 2020), Chat-GLM (Zeng et al., 2024), GPT-4 (Achiam et al., 2023), GPT-3.5 (OpenAI, 2022) and Gemini (Anil et al., 2023). These models demonstrate strong capabilities in logic and mathematics (Imani et al., 2023; Romera-Paredes et al., 2024; Liu et al., 2023; Shi et al., 2023) and scientific domains (Singhal et al., 2023; Bingler et al., 2022), raising the prospect of AI agents that can support scientific discovery through physics reasoning. However, despite these advances, foundation models still face serious limitations: they lack domain-specific expertise (Li et al., 2023a; Shen et al., 2023), frequently hallucinate factual content (Ji et al., 2023; Xiong et al., 2024), and struggle to consistently apply the appropriate physical principles in problem-solving.

Retrieval-Augmented Generation (RAG) has recently emerged as a promising approach to mitigating the limitations of LLMs by integrating external knowledge sources into their workflows (Caffagni et al., 2024; Gao et al., 2023). Yet, its effectiveness for physics reasoning remains largely unexplored. In this context, RAG offers a natural solution: just as a student preparing for an exam consults past competition problems to recall formulas and problem-solving strategies, an LLM equipped with retrieval can access relevant laws, examples, and reasoning patterns from a curated retrieval corpus. This mechanism not only improves factual accuracy but also helps guide the selection and application of physical principles in diverse contexts.

A key obstacle, however, is the absence of suitable benchmarks. Whereas mathematics has benefited from a rich ecosystem of high-quality datasets and benchmarks (Zheng et al., 2022; Hendrycks et al., 2021; Cobbe et al., 2021; Clark et al., 2018; Azerbayev et al., 2024; Fang et al., 2025), physics-specific evaluations are scarce. Existing natural science datasets (Welbl et al., 2017; Lu et al.,

*Equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21927–21956
November 4–9, 2025

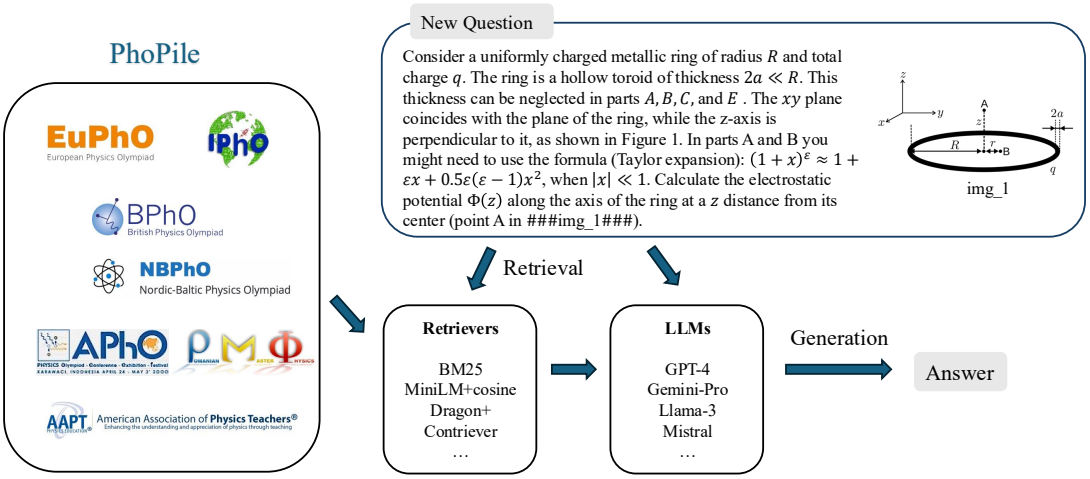

Figure 1: PhoPile and the overall workflow of foundation models with RAG.

2022; Chen et al., 2023) contain only a small number of low-difficulty, text-only physics problems. OlympiadBench (He et al., 2024) raises the level of challenge but evaluates models in isolation, without retrieval. Furthermore, evaluation itself poses unique challenges: physics answers may take diverse forms—numerical, symbolic, or diagrammatic—making automatic grading substantially more difficult than in mathematics.

To address these gaps, we introduce **PhoPile**[1], the first multimodal RAG benchmark for physics Olympiad problems, as illustrated in Figure 1. PhoPile consists of 390 Olympiad-level physics problems from 2019–2021 for evaluation, along with 2,662 problems from earlier years that serve as an external retrieval database. Using PhoPile, we benchmark RAG-augmented foundation models, covering both LLMs and LMMs, together with a variety of retrievers. We highlight two key observations: competition problems share similar concepts across years, and past problems capture not only the necessary physics knowledge from basic to advanced levels, but also valuable strategies for applying laws and formulas to novel scenarios. Importantly, PhoPile incorporates multimodal content—including diagrams, graphs, and equations—mirroring the real-world practice of physics problem solving. We further design an LLM-as-judge framework for evaluation. Our approach uses instructions and reference solutions to automatically score generated outputs, incorporating both step-wise and solution-level assessments to capture the richness of physics reasoning. This framework enables scalable evaluation that is otherwise infeasible with traditional script-based methods.

Our contributions are threefold:

- We propose PhoPile, the first multimodal benchmark for evaluating retrieval-augmented physics reasoning.
- We introduce a new LLM-as-judge evaluation framework tailored for physics, capable of handling diverse solution formats.
- We conduct a comprehensive benchmark of 8 foundation models with 4 text-only retrievers and 3 multimodal retrievers, providing the first systematic study of RAG for physics reasoning.

2 The PhoPile dataset

2.1 Overview

PhOPile is structured to evaluate the performance of the RAG pipeline in the domain of high-level physics problem-solving. Therefore, we collect Physics Olympiad questions from various regions around the globe, including International Physics Olympiad (IPho), Asian Physics Olympiad (APhO), European Physics Olympiad (EuPhO), Nordic-Baltic Physics Olympiad (NBPhO), Romanian Master of Physics (RMPh), United States Physics Olympiad (USAPhO), and British Physics Olympiad (BPhO), all of which are publicly available.

Our data collection is motivated by the real-world practice of examinees reviewing past exam

[1]Data and code available at: https://github.com/aialt/PhoPile

problems when preparing for future tests. Physics competition problems across years often share overlapping knowledge points, cover essential concepts from basic to advanced levels, and showcase diverse strategies for applying physical formulas. Organizing the dataset in this way provides rich references for tackling new problems by leveraging the knowledge and methods embedded in past competitions.

To evaluate foundation models with RAG in the context of physics, we organize the collected data as follows:

- Evaluation Set: 390 Olympiad-difficulty-level physics problems from 2019-2021, used to evaluate the model's performance on contemporary problems;
- Retrieval Corpus: $2,662$ problems before 2019, used by the retriever to provide context and reference for solving new problems.

The evaluation set is further divided into two subsets: *PhoPile-Test*, which contains 267 questions from 125 main problems in text-only form, and *PhoPile(V)-Test*, which consists of 123 questions from 77 main problems that include images either in the question statement or in the reference solutions.

2.2 Data Collection

In this section, we describe the process of collecting and standardizing physics competition problems into a unified format. A typical physics problem consists of textual descriptions, mathematical formulas, and images, and often exhibits hierarchical sub-question structures with multiple reference solutions. All finalized samples are stored in JSON format.

To construct PhoPile in a consistent and high-quality manner, we applied the following preprocessing steps:

(1) Text Cleaning. We delete extraneous elements from the questions, like the 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.

(2) Formula Representation in LaTeX. As physics problems often involve mathematical formulas, we use LaTeX to represent the solutions with detailed information in plain text, inspired by the LaTeX-based representation adopted in MathPile (Wang et al., 2024). Additionally, since the textual formulation of many physics problems is not directly accessible, we utilize the MathPix OCR tool[2] to convert the content of images into LaTeX code, avoiding extensive manual effort while ensuring consistency.

(3) Image Processing. 32% of problems in PhoPile involve images, a detailed breakdown of image usage is provided in Figure 1 and Figure 2; we store them in the local repository and list the local URLs for the images associated with the questions and the solutions in PhoPile, respectively. To highlight the position of the image appearing in the text, we mark n-th image as `###img_n###`. Regarding the captions of these images, similar to MathVista (Lu et al., 2024), we omit the useless image labels (like 'Figure 05' or 'figure 1') and add the meaningful captions that contain crucial content relevant to the problem into the question description, such as 'Figure 1: Isosceles glass prism with an apex angle of $90°$'.

(4) Hierarchical Question Structure. Unlike mathematical datasets, physics problems often contain several sub-questions that may need to be answered in a specific order. We organize these sub-questions using Arabic numerals as indices. For the rest of the paper, we do not distinguish which main question the sub-questions belong to, except during evaluation.

(5) Multiple Solutions. For certain questions, the source files provide multiple solutions. They often appear as 'Solution 2', 'Another way to solve this problem'. For the completeness of the question and subsequent development, we also store them indexed by 'solution 1', 'solution 2', *etc.* Illustrative examples are provided in Appendix.

2.3 Data Analysis

Statistics. Token statistics are summarized in Table 2. With the exception of EuPho, the majority of questions and solutions contain fewer than 500 tokens. This size is well within the context window of current popular LLMs such as Llama-2 (Touvron et al., 2023) and its variants, allowing for complete inference or training without the need for prompt pruning.

Images. The overall image statistics are summarized in Table 1. Images are widely present in PhoPile, appearing in both questions and solutions,

[2] https://mathpix.com

Evaluation Set	
w/ image	117
w/o image	273
Retrieval Corpus	
w/ image	879
w/o image	1,783

Table 1: Number of questions with or without images.

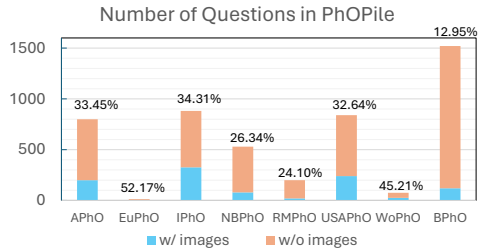

Figure 2: Proportion of image-containing questions across different dataset sources.

Source		# Questions	# Tokens	# Tokens per q (Max/Min/Ave)	Years
APhO	evaluation set	87	19,231	1,677/10/221	2019-2021
	retrieval corpus	502	85,949	1,708/7/171	2000-2018
EuPhO	evaluation set	15	2,489	3,951/20/902	2019-2021
	retrieval corpus	8	1,434	278/24/179	2017-2018
IPhO	evaluation set	93	16,871	854/15/181	2019-2021
	retrieval corpus	854	120,995	846/8/142	1967-2018
NBPhO	evaluation set	55	6,856	468/15/125	2019-2021
	retrieval corpus	374	36,400	569/5/97	2003-2018
RMPh	evaluation set	63	11,828	1,399/15/188	2019-2021
	retrieval corpus	132	15,251	792/9/116	2012-2018
USAPhO	evaluation set	77	6,373	262/10/82	2019-2021
	retrieval corpus	646	58,908	739/4/91	2007-2019
WoPhO	evaluation set	0	-	-	-
	retrieval corpus	146	17,075	813/9/117	2011-2013
Total	evaluation set	390	63,648	3,951/10/	2019-2021
	retrieval corpus	2,662	336,012	1,708/4/	1967 - 2019

Table 2: The token statistics of PhOPile.

with 33% of problems containing at least one image.

Their distribution across different sources is shown in Figure 2. These images play an essential role in conveying information such as experimental setups, physical systems, and data visualizations that cannot be fully captured by text alone.

In our PhoPile, about two-thirds of the solutions include images, ranging from curve plots that illustrate variable relationships to structural diagrams for force analysis, among other types. Although most current LMMs are unable to generate images that precisely meet the requirements posed by the problems, we deliberately preserve these instances to encourage and support future research on multimodal reasoning in physics.

3 Experiments

In this section, we introduce the RAG pipeline, describe the evaluation workflow, and present the experimental results of foundation models.

3.1 RAG Pipeline

The RAG pipeline comprises two main components—the retriever and the generator—and we further incorporate a reflection mechanism to enhance performance.

Retriever. Given an input query q, the retriever searches an external retrieval corpus $\mathcal{D}$ to find the most relevant problems and solutions. A scoring function $f(q, d_i)$ assigns relevance scores to each item $d_i \in \mathcal{D}$, and the top-k items with the highest scores are selected:

$$\mathcal{R}(q) = \{d_1, d_2, \ldots, d_k\} \quad \text{where} \quad d_i \in \mathcal{D}.$$

In PhoPile, $\mathcal{D}$ consists of 2,662 physics competition problems with reference answers collected before 2019. The retriever returns the top-k relevant question–answer pairs $\mathcal{R}(q) = \{(q_i, a_i)\}_{i=1}^{k}$, where q_i and a_i denote the retrieved question and answer, respectively. The scoring function can thus be rewritten as $f(q, q_i)$.

As there is no domain-specific retriever for

Your task is to answer the physics questions. The mathematical formulas are provided in Latex code. There are some related questions and their answers you may find helpful. \n Here are the examples:
Question: {Retrieved Question 1}
Reference answer: {Reference Answer to Question 1}
Question: {Retrieved Question 2}
Reference answer: {Reference Answer to Question 2}\n
The question that you need to solve is: \n {Question to be answered} \n\n
Respond with the FINAL answer to the question to get a higher score as possible as you can, rather than only give directions or suggestions for solving the problem. Do NOT use the conditions in the example questions to solve the question.

Figure 3: Instruction prompt template for the generator to answer the question.

physics, we adopt general-purpose retrieval methods. For text-only retrieval, we adopt several representative methods: (i) a sparse retriever, BM25 (Robertson and Zaragoza, 2009); and (ii) dense retrievers, including Emb-cos (embedding model *all-MiniLM-L6-v2* (Wang et al., 2020) with cosine similarity), Dragon+ (Lin et al., 2023), and Contriever (Izacard et al., 2022). For multimodal retrieval, we employ CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and VisualBERT (Li et al., 2019) to obtain joint text–image embeddings, and use cosine similarity to identify the closest problems, which are likewise used as exemplars.

Generator. The generator takes the retrieved pairs $\mathcal{R}(q)$ together with the original query q and produces a coherent, contextually grounded response r:

$$r = G(p(q, \mathcal{R}(q))),$$

where p denotes the prompt template that integrates the retrieved information, and G is the generative model. An illustration of the prompt design is provided in Figure 3. In our evaluation, we benchmark both text-based LLMs and vision–language LMMs, including Llama-2-13B, GPT-3.5, Llama-3-70B (Grattafiori et al., 2024), DeepSeek-Math (Shao et al., 2024), GPT-4, Gemini-Pro (Georgiev et al., 2024), GPT-4V (OpenAI, 2023), and Gemini-Pro-Vision.

Reflection. We use a reflection mechanism based on GPT-4. While retrieved questions and solutions can improve the generator's performance, they may also introduce noise that misleads reasoning. One cause is the long input context after retrieval, which can dilute focus and increase distraction. The reflection mechanism can mitigate this issue. As illustrated in Figure 4, given a physics problem and two candidate answers—one generated with RAG and one without—the model is prompted to compare their relative quality and select the response it considers more accurate. This self-reflection step reduces the negative impact of noisy retrievals and enhances the robustness of the final output.

Your task is to choose the answer with a higher score of the given physics problem.\n\n
Question: {Question to be answered} \n
Answer 1: {Candidate answer without RAG} \n
Answer 2: {Candidate answer with RAG}\n\n
Please give a reason and output the final answer number in side "##", for example, ##1##.

Figure 4: Instruction prompt template for reflection to choose the answer with or without RAG.

Workflow. Given a problem, we apply our RAG pipeline to solve it. For problems with a single question, we directly prompt the LLM to solve it while inserting the retrieved question and answers into the prompt. For problems with multiple sub-questions, we retrieve question and answer pairs according to the sub-question to be solved and incorporate the retrieved context into the query. The LLM is then prompted to generate a solution, which is appended to the query as additional context. This updated query is fed back to the LLM to solve the next sub-question. The process repeats until all sub-questions under the same main problem are answered, enabling the model to leverage both retrieved knowledge and its own prior reasoning history.

3.2 Evaluation

k	0	1	2	3
Accuracy (%)	37	49	73	87

Table 3: Accuracy of GPT-4 as a grader across different tolerance thresholds k.

Marking Scheme. Physics Olympiad problems typically require long chains of logical reasoning; therefore, scoring must account not only for the correctness of the final answer but also for the quality of intermediate reasoning steps. Instead of

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.
Standard answer: {Reference Answer}
Student answer: {Candidate Answer from Generator}

Figure 5: Prompt template for the evaluator to score the candidate answer, given the reference answer.

adopting the original competition-specific marking schemes, which assign varying maximum scores across problems, we standardize the scoring by setting the maximum score of each problem to ten points.

To this end, we design an *LLM-as-judge* evaluation framework tailored for physics, supporting both holistic and step-wise scoring. As illustrated in Figure 5, GPT-4 is employed as the grader: it is prompted to compare candidate model solutions against the reference solution and assign a score between 0 and 10.

To examine the reliability of GPT-4 as a grader, we conducted a human evaluation on the *PhoPile-Test* set. Three experienced instructors who regularly provide training for the IPhO independently graded the model-generated solutions based on the same rubric that was provided to the LLMs for scoring. The final human score for each solution was obtained by averaging the three annotators' scores. GPT-4 was prompted with the same rubric and asked to assign scores to the same set of solutions. We then compared GPT-4's scores with the averaged human annotations under varying tolerance thresholds k, where k denotes the maximum allowable difference between the GPT-4 score and the human-assigned score, as shown in Table 3. Results show that although exact agreement is not achieved, GPT-4 provides sufficiently consistent judgments to capture relative performance differences across models.

Evaluation Metric. We report both the average score (AS) the LLMs and LMMs earn and their pass rate (PR) over *PhoPile-Test* and *PhoPile(V)-Test*. The average score is defined as,

$$\textbf{AS} = \frac{\text{Total points gained by candidate model}}{\text{Number of questions} \times 10} \times 100\%. \quad (1)$$

We regard the generator (LLMs and LMMs) as successfully passing a problem if they answer the problem correctly and earn a score of 10. Therefore pass rate is defined as,

$$\textbf{PR} = 100\% \cdot \frac{N_{\text{full-score}}}{N_{\text{total}}}. \quad (2)$$

Intuitively, the pass rate and average score reflect the performance of the retriever and generator in a coupled manner. The higher the value of the pass rate and average score, the better the performance of the retriever and generator.

Baselines. For the *PhoPile-Test* subset containing text-only questions, we evaluate strong publicly available instruction-tuned LLMs, including LLaMA-3, as well as proprietary models trained with private data, such as GPT-3.5, GPT-4, and Gemini-Pro. We additionally evaluate DeepSeekMath-7B, which is trained with math-related tokens and achieves comparable performance of GPT-4.

For the *PhoPile(V)-Test* subset with images, we use LMMs including GPT-4V and Gemini-Pro-Vision. By default, we evaluate the zero-shot CoT (Wei et al., 2022) performance for the baselines. To evaluate RAG's effectiveness in solving Olympic-level mathematical physics problems, we experiment with the different RAG methods, where a model generates output given the input question prepended with the top-K retrieved examples.

In addition, we finetune several open-source models on the retrieval corpus, which consists of question–answer pairs. The models include Mistral-7B-v0.3, Phi-3.5-mini, LLaMA-3-8B, and Mathstral-7B-v0.1. We denote the fine-tuned versions with the suffix '-FT', for example: Mistral-7B-v0.3-FT, Phi-3.5-mini-FT, LLaMA-3-8B-FT, and Mathstral-7B-v0.1-FT.

3.3 Main Results

We report the evaluation results of the foundation models with text-only retrieval under the 2-shot setting in Table 4, including LLMs on *PhoPile-Test* and LMMs on *PhoPile(V)-Test*. RAG provides some insights to aid the generators' physical reasoning. For example, Gemini-Pro combined with Contriever improves substantially, from 17.18%

Model	Input	w/o RAG PR(AS)	Emb-cos PR(AS)	BM25 PR(AS)	Dragon+ PR(AS)	Contriever PR(AS)
On *PhoPile-Test*						
Llama-3-70B	*T*	10.51(1.34)	5.4(1.84)	19.07(4.86)	13.62(4.83)	10.28(4.65)
Llama-3-70B w/ Reflection	*T*	10.51(1.34)	19.38(4.35)	19.38(4.35)	14.51(4.81)	10.80(4.60)
DeepSeek-Math	*T*	4.10(0.64)	2.06 (0.27)	2.06(0.29)	2.06(5.93)	3.08(0.38)
DeepSeek-Math w/ Reflection	*T*	4.10(0.64)	16.95(2.85)	3.59 (0.55)	2.83(6.085)	3.37 (0.54)
GPT-3.5	*T*	7.95(4.12)	8.72(4.02)	8.23(3.84)	10(3.75)	7.69(3.91)
GPT-3.5 w/ Reflection	*T*	7.95(4.12)	11.79(4.9)	9.23(4.37)	10.26(4.25)	8.46(43.21)
Gemini-Pro	*T*	17.18(5.30)	16.15(4.91)	15.90(4.93)	16.41(5.69)	**30.51(5.19)**
Gemini-Pro w/ Reflection	*T*	17.18(5.30)	21.54(5.72)	20.51(5.56)	18.72(5.65)	19.74(5.49)
GPT-4	*T*	26.41(6.27)	24.10(5.71)	25.19(5.92)	25.71(5.91)	25.19(5.82)
GPT-4 w/ Reflection	*T*	26.41(6.27)	27.92(6.22)	27.69(6.37)	28.46(6.34)	26.99(6.23)
Mistral-7B-v0.3-FT	*T*	1.47(2.10)	22.64(5.38)	20.90(4.10)	25.28(4.62)	23.28(6.15)
Phi-3.5-mini-FT	*T*	2.56(1.95)	18.00(7.18)	16.25(6.43)	20.31(7.95)	21.44(7.46)
Llama-3-8B-FT	*T*	5.86(2.17)	28.31(5.90)	26.44(5.38)	27.46(5.91)	25.39(6.19)
Mathstral-7B-v0.1-FT	*T*	6.62(2.84)	27.17(5.91)	**29.02(9.28)**	28.90(8.21)	27.66(8.74)
On *PhoPile(V)-Test*						
Gemini-Pro-V	*T, I*	12.82(5.09)	17.95(5.24)	12.82(4.78)	12.88(4.858)	14.96(5.04)
Gemini-Pro-V w/ Reflection	*T, I*	12.82(5.09)	19.23(5.24)	16.67(5.05)	15.38(4.83)	17.09(5.06)
GPT-4V	*T, I*	21.79(6.26)	20.51(5.43)	7.69(2.50)	21.46(5.53)	21.79(5.65)
GPT-4V w/ Reflection	*T, I*	21.79(6.26)	**21.89(6.09)**	19.31(5.25)	21.03(6.16)	21.46(6.20)

Table 4: Evaluation results on PhoPile. *Input: T* = question text only; *T, I* = question text with images. Values indicate pass rates (PR) in percentages, with average scores (AS) in parentheses. Bold values denote the best performance. All retrievers are text-only.

Model	Input	w/o RAG PR(AS)	CLIP PR(AS)	VisualBERT PR(AS)	ALIGN PR(AS)
Gemini-Pro-V	*T, I*	12.82(5.09)	17.48(4.99)	13.59(3.42)	14.56(5.88)
Gemini-Pro-V w/ Reflection	*T, I*	12.82(5.09)	14.56(5.12)	**17.48(5.28)**	15.53(5.35)
GPT-4V	*T, I*	21.79(6.26)	**30.10(6.20)**	24.27(5.80)	15.53(5.79)
GPT-4V w/ Reflection	*T, I*	21.79(6.26)	26.41(5.99)	22.33(5.58)	23.30(5.71)

Table 5: Evaluation results on *PhoPile(V)-Test* with multimodal retrieval. Bold values denote the best-performing retriever for each model.

to 30.51%, while LLaMA-3-70B with BM25 increases from 10.51% to 19.07%. However, not all retrievers yield positive effects. In many cases, performance decreases with RAG, primarily due to noise and irrelevant content introduced by the retrieved examples. For LMMs with RAG, GPT-4 consistently outperforms Gemini in both pass rate and average score. With the incorporation of RAG, Gemini-Pro improves from 12.82% to 17.95%. We also investigate reflection in the RAG framework as a means to mitigate the negative impact of retrieved questions. This mechanism yields noticeable performance improvements.

Interestingly, among open-source models, we observe substantial improvements after fine-tuning even for those with fewer than 8 billion parameters. Their overall performance increased by factors ranging from 5 to 17. The strongest model, Mathstral, achieved an accuracy of 29.02, which is already comparable to the best closed-source result of 30.51.

The results of multimodal retrieval based on joint text–image embeddings are presented in Table 5. Both Gemini-Pro-V and GPT-4V show improvements after applying multimodal RAG, though the extent varies across retrievers. GPT-4V benefits most from CLIP, reaching a pass rate of 30.10%, while Gemini-Pro-V gains more from VisualBERT, where reflection boosts its performance to 17.48%. These results demonstrate that the choice of multimodal retriever significantly impacts performance, with CLIP particularly effective for GPT-4V and VisualBERT more favorable for Gemini-Pro-V.

Error analysis. In many cases, retrieved examples had a negative impact on performance, primarily due to the following reasons: 1) The gen-

Model	#Shots	Emb-cos PR(AS)	BM25 PR(AS)	Dragon+ PR(AS)	Contriever PR(AS)
GPT-3.5	1	8.97(3.88)	6.92(3.65)	9.74(3.87)	0.77(0.62)
	2	8.72(4.02)	8.23(3.84)	10.00(3.75)	7.69(3.91)
	3	9.74(3.90)	6.41(3.77)	7.44(3.70)	7.71(3.88)
GPT-4	1	26.74(6.00)	22.82(5.70)	26.41(6.01)	28.97(6.10)
	2	24.10(5.71)	25.19(5.92)	25.71(5.91)	25.19(5.82)
	3	25.90(6.01)	22.56(5.65)	22.37(5.89)	24.62(5.91)

Table 6: Evaluation results of GPT-3.5 and GPT-4 with different numbers of retrieved examples. The rows correspond to different numbers of shots (1, 2, and 3), with values in parentheses indicating standard deviations.

eral retriever was not effectively applied to physics problems, as retriever specific to physics may consider the questions that using the same theorem as the top-k relevant ones, instead of those with highest semantic similarity. Therefore, it highlights the significance of establishing domain-specific retrievers. 2) The format in retrieved questions misleads the candidate models' answering. The retrieved questions and their reference answer may provide guidance answers instead of directly answering the question. Therefore, the foundation models may refuse to answer the final answer directly and answer with some guidance for the question. 3) Additionally, some wrong answers arise from using conditions in the retrieved questions as if they were the known conditions in the current question, demonstrating the significance of noise robustness in foundation models with RAG. Please refer to Appendix for the examples.

3.4 Ablation

Table 6 reports the evaluation results of GPT-3.5 and GPT-4 using different retrieval methods: Emb-cos, BM25, Dragon+, and Contriever, with varying numbers of retrieved examples: 1-shot, 2-shot, and 3-shot. The results show that the top-1 retrieved example from Contriever can negatively affect the performance of GPT-3.5, whereas Emb-cos and Dragon+ provide more useful examples, particularly in the 1-shot and 2-shot settings.

4 Related Work

RAG (Chen et al., 2024; Zhang et al., 2024) has recently drawn significant attention in complimenting the domain-specific expertise for LLMs (Zhang et al., 2025b), or constructing demonstrations for in-context learning (ICL) (Poesia et al., 2022; Agrawal et al., 2023; Liu et al., 2022; Hu et al., 2022; Li et al., 2023b), thus serves as a natural way to enhance foundation model's capability of physic reasoning by integrating external knowledge sources. Let's consider a high school student preparing for an exam; it is natural for the student to review past exam questions to find similar types of problems to practice. These questions can provide similar problem-solving approaches and relevant knowledge applications, much like how RAG can retrieve and incorporate pertinent information and similar demonstrations to enhance the reasoning and accuracy of LLMs, resulting in more informed and contextually relevant responses.

In recent years, LLMs have developed very rapidly, providing great convenience for people's needs in all aspects of life (Wang et al., 2025; Liu et al., 2024; Hu et al., 2025; Fang et al., 2024; Schipper et al., 2025). These models, like GPT-3, GPT-4 and Gemini have already shown great performance in terms of accuracy, interpretability, and multimodality, similarly as general LLMs, they show outstandingly high performance of natural science QA and mathematical reasoning. Meanwhile, a range of excellent open source models, including T5 (Raffel et al., 2020), GPT-2 (Radford et al., 2019) and Llama-2, is available for researchers to enhance further, by training them on a specialized dataset to attain superior capabilities compared to generalized models (Magister et al., 2023; Shridhar et al., 2023; Zhang et al., 2025a). Consequently, a series of outstanding open source models that are specifically trained and fine-tuned on math have emerged, such as DeepSeekMath, Llema and Goat (Liu and Low, 2023). Additionally, there are also a few models focusing on formal proof such as LeanDojo (Yang et al., 2023); 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., 2024) and LeanDojo 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.

Models which demonstrate excellent performance on mathematical ability are inseparable from high-quality datasets and corpus such as Mathpile, proof-pile-2, MiniF2F, MATH, GSM8K, MLFMF (Bauer et al., 2023), MathOdyssey and the corpus proposed by DeepSeekMath. 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 introduced a specialized image-based mathematical dataset and conducted evaluations of models such as GPT-4, GPT-3.5, 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, ScienceQA, C-eval (Huang et al., 2023), E-EVAL (Hou et al., 2024), and TheoremQA.

5 Conclusion

In this work, we present **PhoPile**, a benchmark designed to comprehensively evaluate the ability of foundation models to perform physics reasoning with retrieval-augmented generation (RAG) across both text-only and image-based questions. We benchmark a range of mainstream foundation models, including both large language models (LLMs) and large multimodal models (LMMs), together with multiple retrievers. To ensure robust evaluation, we introduce an LLM-as-judge framework capable of assessing diverse solution formats. Our results demonstrate that combining physics corpora with retrieval can improve performance, while also revealing challenges that motivate further research in retrieval-augmented physics reasoning.

Limitations

Our retrieval corpus is limited in scale due to constraints in data acquisition, including restricted access to diverse sources and practical challenges in collection and integration. As future work, we plan to incorporate multimodal cross-referencing, which would enable richer interactions between text, formulas, and images, and has the potential to further improve the accuracy and robustness of foundation models for physics reasoning.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Anthropic. 2023. Claude 2. https://www.anthropic.com/news/claude-2. Accessed: 2024-02-14.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Qiaochu Jiang, Jia Deng, Stella R Biderman, and Sean Welleck. 2024. Llemma: An open language model for mathematics. In *International Conference on Learning Representations*, volume 2024, pages 40622–40649.

Andrej Bauer, Matej Petković, and Ljupco Todorovski. 2023. Mlfmf: Data sets for machine learning for mathematical formalization. In *Advances in Neural Information Processing Systems*, volume 36, pages 50730–50741. Curran Associates, Inc.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2022. Cheap talk and cherry-picking: What climatebert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47:102776.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1818–1826.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. TheoremQA: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901, Singapore. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

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

Meng Fang, Shilong Deng, Yudi Zhang, Zijing Shi, Ling Chen, Mykola Pechenizkiy, and Jun Wang. 2024. Large language models are neurosymbolic reasoners. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Meng Fang, Xiangpeng Wan, Fei Lu, Fei Xing, and Kai Zou. 2025. Mathodyssey: Benchmarking mathematical problem-solving skills in large language models using odyssey math data. *Scientific data*, 12(1):1392.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and others. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1):32.

Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

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. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7753–7774, Bangkok, Thailand. Association for Computational Linguistics.

Xueyu Hu, Tao Xiong, Biao Yi, Zishu Wei, Ruixuan Xiao, Yurun Chen, Jiasheng Ye, Meiling Tao, Xiangxin Zhou, Ziyu Zhao, and others. 2025. OS agents: A survey on MLLM-based agents for computer, phone and browser use. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7436–7465, Vienna, Austria. Association for Computational Linguistics.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A. Smith, and Mari Ostendorf. 2022. In-context learning for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2627–2643, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xijie Huang, Li Lyna Zhang, Kwang-Ting Cheng, Fan Yang, and Mao Yang. 2024. Fewer is more: Boosting math reasoning with reinforced context pruning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13674–13695, Miami, Florida, USA. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, and others. 2023. C-eval: a multi-level multi-discipline chinese evaluation suite for foundation models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. MathPrompter: Mathematical reasoning using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational*

Linguistics (Volume 5: Industry Track), pages 37–42, Toronto, Canada. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12).

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023a. Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? a study on several typical tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 408–422, Singapore. Association for Computational Linguistics.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023b. Unified demonstration retriever for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4644–4668, Toronto, Canada. Association for Computational Linguistics.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6385–6400, Singapore. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Tiedong Liu and Bryan Kian Hsiang Low. 2023. Goat: Fine-tuned llama outperforms gpt-4 on arithmetic tasks. *arXiv preprint arXiv:2305.14201*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, and others. 2024. Agentbench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*.

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 preprint arXiv: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*.

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 *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2022. Introducing chatgpt. Accessed: 2023-02-06.

OpenAI. 2023. Gpt-4v(ision) system card.

Gabriel Poesia, Alex Polozov, Vu Le, Ashish Tiwari, Gustavo Soares, Christopher Meek, and Sumit Gulwani. 2022. Synchromesh: Reliable code generation from pre-trained language models. In *International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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

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.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Bernardino Romera-Paredes, Mohammadreza Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Demis Hassabis. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475.

Olivier Schipper, Yudi Zhang, Yali Du, Mykola Pechenizkiy, and Meng Fang. 2025. Pillagerbench: Benchmarking llm-based agents in competitive minecraft team environments. In *2025 IEEE Conference on Games (CoG)*, pages 1–15.

Raymond A. Serway, Robert J. Beichner, and John W. Jewett. 2000. *Physics for Scientists and Engineers*. Saunders College Publishing.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. In chatgpt we trust? measuring and characterizing the reliability of chatgpt. *arXiv preprint arXiv:2304.08979*.

Zijing Shi, Meng Fang, Shunfeng Zheng, Shilong Deng, Ling Chen, and Yali Du. 2023. Cooperation on the fly: Exploring language agents for ad hoc teamwork in the avalon game. *arXiv preprint arXiv:2312.17515*.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073, Toronto, Canada. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and others. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

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

Jiaqi Wang, Enze Shi, Huawen Hu, Chong Ma, Yiheng Liu, Xuhui Wang, Yincheng Yao, Xuan Liu, Bao Ge, and Shu Zhang. 2025. Large language models for robotics: Opportunities, challenges, and perspectives. *Journal of Automation and Intelligence*, 4(1):52–64.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc.

Zengzhi Wang, Xuefeng Li, Rui Xia, and Pengfei Liu. 2024. Mathpile: A billion-token-scale pretraining corpus for math. In *Advances in Neural Information Processing Systems*, volume 37, pages 25426–25468. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

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*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, and others. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Yudi Zhang, Lu Wang, Meng Fang, Yali Du, Chenghua Huang, Jun Wang, Qingwei Lin, Mykola Pechenizkiy, Dongmei Zhang, Saravan Rajmohan, and others. 2025a. Distill not only data but also rewards: Can smaller language models surpass larger ones? *arXiv preprint arXiv:2502.19557*.

Yudi Zhang, Pei Xiao, Lu Wang, Chaoyun Zhang, Meng Fang, Yali Du, Yevgeniy Puzyrev, Randolph Yao, Si Qin, Qingwei Lin, and others. 2025b. Ruag: Learned-rule-augmented generation for large language models. In *The Thirteenth International Conference on Learning Representations*.

Zihan Zhang, Meng Fang, and Ling Chen. 2024. RetrievalQA: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6963–6975, Bangkok, Thailand. Association for Computational Linguistics.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. 2022. minif2f: a cross-system benchmark for formal olympiad-level mathematics. In *International Conference on Learning Representations*.

Appendix

1. **Question Examples in PhOPile** (Sec. A)
   Dataset indexing scheme ("question_number", "sub_question_number", "sub_sub_question_number") and textual normalization of in-question indices.
   Figure: Examples from the dataset (Figure 6).

2. **Supplementary Experiments and Examples for GPT-4 Scoring** (Sec. B)
   Scoring rubric and qualitative marking cases for GPT-4 as evaluator.
   Table: Candidate-answer categories and expected scores (Table 7).
   Figures: 10/5/0-score answer exemplars (Figures 7–9).

3. **Prompt Examples** (Sec. C)
   Prompt formats for pure-text inference and 1-shot RAG.
   Figures: Pure text prompt (Figure 10); 1-shot RAG prompt (Figure 11).

4. **Gemini Example** (Sec. D)
   One worked Gemini inference case with 1-shot RAG.
   Figure: Gemini output example (Figure 12).

5. **Multimodality Example** (Sec. E)
   Vision-augmented physics problem; standard answer; GPT-4-V vs Gemini-Pro-Vision responses and scores.
   Figure: The image from the problem (Figure 13).

6. **Runtime Analysis** (Sec. F)
   Retrieval vs retrieval+embedding timing across methods.
   Table: RAG performance (Table 8).

7. **Examples of Answer Errors** (Sec. G)

   (a) *Error Type 1:* Gives guidelines instead of a direct answer (projectile problem).
   (b) *Error Type 2:* Misguided by retrieved examples (thermoacoustic temperature amplitude).
   (c) *Error Type 3:* Wrong retrieval (dipole torque/power vs unrelated contexts).

8. **Additional Results** (Sec. H)
   Difficulty distribution; RAG-driven zero $\rightarrow$ non-zero improvements; average scores by difficulty.
   Tables: Difficulty distribution (Table 9); average scores by difficulty (Table 10); questions improved from 0 to non-0 after RAG (Table 11).

A 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`'. We provide examples in Figure 6.

B Supplementary Experiments and Examples for GPT-4 Scoring

As shown in Table 7, we verify the GPT4's capability on marking the candidate answers by foundation models, given the reference solution.

Figure 7, Figure 8 and Figure 9 are examples of evaluator marking three different types of answers.

Example 1:
“index”: 2882,
“problem”: "A metal ring of radius R is made out of a wire of cross section aria s; the resistivity of the wire’s material is given by ρ. The ring is place in a vertical plane. A small magnetic needle is placed in the center of the ring. The magnetic needle can rotate freely around the vertical axis which passes through the center of the ring. Due to mechanical inertia, the needle points in the direction of the average horizontal magnetic field at the center of the ring. When the ring is motionless in the reference frame of the Earth, the magnetic needle indicates the direction of the horizontal component of Earth’s magnetic field. When the ring is rotating around its vertical axis with a constant angular velocity, ω, the magnetic needle deviates, in horizontal plane, from this direction by an angle α. 1.a. Find the expression of angle of deviation, α of the magnetic needle. Write your expression for α as a function of s, ω, ρ and of the magnetic permeability μ_0.”,
“question_number”: 978,
“sub_question_number”: 1,
“sub_sub_question_number”: null,
“source”: "RMPh”,
“year”: 2021,
“solution”: “For: $\vec{B}_{\text{Earth}} = B_h \cdot \hat{e}_x + B_v \cdot \hat{e}_z$ B_h - horizontal component of Earth’s magnetic field B_v - vertical component of Earth’s magnetic field formula of surface vector $\vec{S}$ for the ring $\vec{S} = \pi \cdot R^2 \cdot (\cos(\omega \cdot t) \cdot \hat{e}_x + \sin(\omega \cdot t) \cdot \hat{e}_y)$ expression of magnetic flux ϕ of Earth’s magnetic field through ring’s surface ...”,
“imgQ”: null,
“imgA”: null

Example 2:
“index”: 2883,
“problem”: “A thin ring of mass m, radius r_0 and inductance L is maintained in a horizontal plane above a cylindrical magnetic bar which is placed vertically (see ###img_1###). The vertical axis of symmetry of the cylindrical magnetic bar is aligned with the center of the ring. The magnetic field due to the cylindrical magnet is shown in ###img_1### and each of its components are given by: the radial component is given by $B_r = B_0 \cdot \beta \cdot r$, while the vertical component is given by $B_z = B_0 \cdot (1 - \alpha \cdot z)$.$B_0, \alpha, \beta$ are all positive constant with appropriate dimensions, while z and r denotes the vertical and, respectively, the radial coordinate of the system. Initially, there is no electric current passing through the ring and it is kept fixed above the magnet. It is then allowed to fall due to the gravitational pull of the Earth, given by the gravitational acceleration, g. During the fall the ring will still be in a horizontal plane and will have the same vertical axis. Answer the following questions and write your results as a function of the variables specified above. Derive the equations of motion for the ring in the reference frame specified in ###img_1###.”,
“question_number”: 978,
“sub_question_number”: 2,
“sub_sub_question_number”: 1,
“source”: "RMPh”,
“year”: 2021,
“solution”: “For: expression of magnetic flux through ring’s surface $\Phi = B_z \cdot \pi \cdot r_0^2 + L \cdot I$ $0 = R \cdot I = \frac{d\Phi}{dt}$ - voltage drop on superconducting ring is zero - magnetic flux inside the ring is constant $\Phi = B_0 \cdot (1 - \alpha \cdot z) \cdot \pi \cdot r_0^2 + L \cdot I =$ constant Initial conditions $\begin{cases} z(t=0) = 0 \\ l(t=0) = 0 \end{cases}$ constant $= B_0 \cdot \pi \cdot r_0^2$ expression of the intensity of electric current through the ring $I = \frac{B_0}{L} \cdot \alpha \cdot \pi \cdot r_0^2 \cdot z$ radial component of the force of interaction is zero - because of symmetry vertical component of the force of interaction $F_z = -\frac{2\pi^2 \cdot \alpha \cdot \beta \cdot B_0^2 \cdot r_0^4}{L} \cdot z$ elastic constant $k = \frac{2\pi^2 \cdot \alpha \cdot \beta \cdot B_0^2 \cdot r_0^4}{L}$ equations of motion for the ring $m \cdot \frac{d^2 z}{dt^2} + k \cdot z = -m \cdot g$ general solution of the equations of motion for the ring ...”,
“imgQ”: ["18.png"],
“imgA”: null

Figure 6: Examples from the dataset.

Candidate Answer	Output
Fully correct solution	10
Fully incorrect solution but with the correct final answer	0
A completely wrong solution	0
Half of the fully correct solution	1-9
Fully correct solution but with a wrong final answer	9
Correct answers in various type (normal format, Latex format, or without units)	10

Table 7: By inputting the candidate solution to be marked and reference answer to GPT4, we can have detailed and based-on-step scores without human intervention.

Example Answer Provided by GPT-3.5 with Cosine Similarity

Input:

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.
Reference answer:
The width of the potential well (L) should be equal to the half of the wavelength of the de Broglie standing wave $\lambda_{\mathrm{dB}}=h / p$, here h is the Planck's constant and p is the momentum of the particle. Thus $p=h / \lambda_{\mathrm{dB}}=h /(2 L)$, and the minimal possible energy of the particle is $E_{\min }=\frac{p^2}{2 m}=\frac{h^2}{8 m L^2}$.
Student answer:
To determine the minimal possible energy \\(E_{\\min}\\) of the quantum particle in the well, we will consider the particle as a standing wave with nodes at the walls of the potential well. This standing wave satisfies the boundary conditions set by the potential well.\n\nIn a one-dimensional potential well, the energy of the particle is quantized. The allowed energy levels are given by:\n\n\\[E_n = \\frac{n^2 h^2}{8 m L^2}, \\]\n\nwhere \\(n \\) is a positive integer representing the quantum energy level, \\(m \\) is the mass of the particle, \\(L \\) is the width of the potential well, and \\(h \\) is the Planck constant.\n\nThe minimal possible energy \\(E_{\\min} \\) of the quantum particle in the well corresponds to the ground state, which is when \\(n = 1 \\). Therefore, substituting \\(n = 1 \\) into the equation above, we get:\n\n\\[E_{\\min} = \\frac{h^2}{8 m L^2}. \\]\n\nThus, the minimal possible energy \\(E_{\\min} \\) of the quantum particle in the well is given by:\n\n\\[\\boxed{E_{\\min} = \\frac{h^2}{8 m L^2}}. \\]

Evaluator's answer:

Assistant: 10

Figure 7: A 10 score example answer provided by GPT-3.5 with cosine similarity.

Example Answer Provided by GPT-3.5 with Cosine Similarity

Input:

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.
Reference answer:
Let us assume $K=k \varepsilon_0^\alpha h^\beta \lambda^\gamma d^\delta$. The SI units of therelevant quantities are: $\left[\varepsilon_0\right]=\frac{\mathrm{A}^2 \cdot \mathrm{s}^4}{\mathrm{~kg} \cdot \mathrm{m}^3}, \quad[h]=\frac{\mathrm{kg} \cdot \mathrm{m}^2}{\mathrm{~s}}, \quad[\lambda]=\mathrm{m}, \quad[d]=\mathrm{A} \cdot \mathrm{s} \cdot \mathrm{m}, \quad[K]=\mathrm{s}^{-1}$. By plugging these expressions into Eq. 2 we obtain a simple system of linear equations for the unknown powers α, β, γ, and δ: $2 \alpha+\delta=0, \quad-\alpha+\beta=0, \quad 4 \alpha-\beta+\delta=-1, \quad-3 \alpha+2 \beta+\gamma+\delta=0$.By solving this system we get: $\alpha=\beta=-1, \quad \gamma=-3, \quad \delta=2$, so that the rate of spontaneous emission is $K=\frac{16 \pi^3}{3} \frac{d^2}{\varepsilon_0 h \lambda^3}$.
Student answer:
To determine the rate of spontaneous emission K in terms of $\\varepsilon_0$, h, $\\lambda$, and d, we can use dimensional analysis. \n\nThe rate of spontaneous emission should have the dimensions of frequency, which is the reciprocal of time. We can express K as being proportional to some combination of the given parameters raised to certain powers. \n\nThe relevant parameters and their dimensions are:\n- $\\varepsilon_0$, with dimensions $[\\varepsilon_0] = \\text{C}^2 \\, \\text{N}^{-1} \\, \\text{m}^{-2}$\n- h, with dimensions $[h] = \\text{J} \\cdot \\text{s}$\n- $\\lambda$, with dimensions $[\\lambda] = \\text{m}$\n- $d \\approx e l$, with dimensions $[d] = \\text{C} \\, \\text{m}$\n\nWe need to combine these parameters in such a way that the dimensions of the expression yield frequency, i.e., $\\text{T}^{-1}$. The rate of spontaneous emission K can be written as:\n$$\nK \\propto \\left(\\frac{1}{\\varepsilon_0 h} \\right)^a \\lambda^b d^c\n$$\n\nBy setting the dimensions of K equal to $\\text{T}^{-1}$, we can determine the powers a, b, and c:\n$$\n\\begin{aligned}\n[\\text{T}]^{-1} & = \\left(\\frac{\\text{C}^2 \\cdot \\text{N}^{-1} \\cdot \\text{m}^{-2}}{\\text{J} \\cdot \\text{s}} \\right)^a \\cdot \\text{m}^b \\cdot \\left(\\text{C} \\cdot \\text{m} \\right)^c \\\\\n\\text{T}^{-1} & = \\text{J}^{-a} \\cdot \\text{C}^{2a+c} \\cdot \\text{N}^{-a} \\cdot \\text{m}^{-2a+b+c} \n\\end{aligned}\n$$\n\nEquating the dimensions on both sides, we have the following system of equations:\n$$\n\\begin{cases}\n-1 = -a & \\text{(equating powers of J)} \\\\\n0 = 2a + c & \\text{(equating powers of C)} \\\\\n0 = -a & \\text{(equating powers of N)} \\\\\n-1 = -2a + b + c & \\text{(equating powers of m)}\n\\end{cases}\n$$\n\nSolving this system of equations will give us the values of a, b, and c. Substituting these values back into the expression for K, along with the given numerical prefactor, we can obtain the complete expression for the rate of spontaneous emission in terms of $\\varepsilon_0$, h, $\\lambda$, and d.

Evaluator's answer:

Assistant: 5

Figure 8: A 5 score example answer provided by GPT-3.5 with cosine similarity.

Example Answer Provided by GPT-3.5 with Cosine Similarity

Input:

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.
Reference answer:
Comparing $f=-E /\left(e C V_0 d \beta\right)$ with $f=-E /(e q d \beta)$ from part C we immediataly obtain $q_{\mathrm{eff}}=C V_0$.
Student answer:
To find the expression for q_{eff} in terms of quantities given in the problem formulation, we need to consider the analogy between the given problem and the reference answers provided.
In the reference answer 1, we have the expression for the total number of modes as N. This total number of modes is related to the parameter q_{eff} in the problem being solved.
In the reference answer 2, the parameters h_2 and r_2 in terms of b and T can be expressed as functions of known quantities. There is an equation relating h_2 to other quantities in a similar way that q_{eff} would be related to known quantities in the current problem.
Therefore, by extending this analogy, we can conclude that the effective quantity q_{eff} in this problem can be expressed as a function of the given quantities b and T in a way similar to how h_2 and r_2 are expressed in reference answer 2. The specific form of this expression will depend on the exact relationships between the different parameters in the problem at hand.

Evaluator's answer:

Assistant: 0

Figure 9: A 0 score example answer provided by GPT-3.5 with cosine similarity.

C Prompt Examples

In this section, we present the prompt examples used to guide LLMs in solving physical questions and evaluating solutions. Figure 10 shows a prompt example of pure text inference, while Figure 11 illustrates a prompt with 1-shot RAG.

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 10: Prompt example of pure text inference

Prompt example for LLMs inference with 1 shot RAG

Input:

Your task is to answer the physics questions. The mathematical formulas are provided in Latex code. There are some related questions and their answers you may find helpful.\nHere are the examples:\n**Question**: In his famous work on magnetic monopoles Paul Dirac has argued that if just one magnetic monopole existed in the Universe, all electric charges should be multiple of a specific elementary electric charge, whose magnitude is related to the magnetic charge of that monopole. Historically, it is the first hypothesis in physics, which explains the existence of the elementary electric charge. Consider the system described in Question 6, assuming that all magnetic monopoles existing in the Nature have magnetic charges of the same magnitude, $+q_m$ and $-q_m$ respectively. By applying the concepts of quantum physics to the motion of electrically charged particle along the circular orbit, derive a relationship between the elementary electric charge e, assumed to be the charge of the electron, and the magnetic charge q_m of the monopole. Calculate q_m numerically.\n**Reference answer**: According to the Bohr's quantization rule, the change of the angular momentum along a specified axis is a multiple of the reduced Plank's constant: $ \\Delta L_z=n \\hbar=n \\frac{h}{2 \\pi} $ By utilizing the expression for $\\Delta L_z$, derived in the last question, we obtain: $ \\frac{\\mu_0 q_m q_e}{2 \\pi}=\\frac{n h}{2 \\pi} \\quad \\text { or } \\quad q_e=n \\frac{h}{\\mu_0 q_m} . $ This result implies that the electric charge is multiple of a universal elementary electric charge $\\frac{h}{\\mu_0 q_m}$. If we identify the elementary electric charge with the charge of the electron, we can find the magnitude of the magnetic charge of the monopole: $ q_m=\\frac{h}{\\mu_0 e}=3.29 \\times 10^{-9} \\mathrm{~A} \\cdot \\mathrm{m} $\n\n**The question that you need to solve is**:\nWave-particle duality, which states that each particle can be described as a wave and vice versa, is one of the central concepts of quantum mechanics. In this problem, we will rely on this notion and just a few other basic assumptions to explore a selection of quantum phenomena covering the two distinct types of particles of the microworld-fermions and bosons.Consider a particle of mass m moving in a one-dimensional potential well, where its potential energy $V(x)$ is given by $V(x)= \\begin{cases}0, & 0 \\leq x \\leq L \\ \\infty, & x<0 \\text { or } x>L\\end{cases}$.While classical particle can move in such a potential having any kinetic energy, for quantum particle only some specific positive discrete energy levels are allowed. In any such allowed state, the particle can be described as a standing de Broglie wave with nodes at the walls.Determine the minimal possible energy $E_{\\min }$ of the quantum particle in the well. Express your answer in terms of m, L, and the Planck's constant h.\n\nResponse with the FINAL answer to the question to get a higher score as possible as you can, rather than only give directions or suggestions for solving the problem. Do NOT use the conditions in the example questions to solve the question.

Output:

Assistant: To determine the minimal possible energy $E_{\\min}$ of the quantum particle in the well, we can use the time-independent Schrödinger equation:\n\n$$-\\frac{\\hbar^2}{2m}\\frac{d^2\\psi(x)}{dx^2} + V(x)\\psi(x) = E\\psi(x)$$\n\nSince the potential energy $V(x)$ is defined to be zero inside the well and infinite outside it, the Schrödinger equation simplifies to:\n\n$$-\\frac{\\hbar^2}{2m}\\frac{d^2\\psi(x)}{dx^2} = E\\psi(x)$$ for $0 \\leq x \\leq L$\n\nThe general solution to this differential equation inside the well is a sinusoidal function:\n\n$$\\psi(x) = A\\sin(kx) + B\\cos(kx)$$\n\nwhere $k = \\sqrt{\\frac{2mE}{\\hbar^2}}$.\n\nGiven that the wavefunction $\\psi(x)$ must go to zero at $x = 0$ and $x = L$ due to the nodes at the walls, we get the boundary conditions: $\\psi(0) = \\psi(L) = 0$. \n\nApplying the boundary conditions, we find that $kL = n\\pi$, where n is a positive integer.\n\nThe allowed energies E_n for the quantum particle are then given by:\n\n$$E_n = \\frac{n^2\\pi^2\\hbar^2}{2mL^2}$$\n\nThe minimal energy $E_{\\min}$ corresponds to $n = 1$, hence:\n\n$$E_{\\min} = \\frac{\\pi^2\\hbar^2}{2mL^2}$$\n\nTherefore, the minimal possible energy $E_{\\min}$ of the quantum particle in the well is $\\frac{\\pi^2\\hbar^2}{2mL^2}$.

Figure 11: Prompt example of pure text inference with 1-shot RAG

D Gemini Example

Figure 12 is an example of an answer provided by Gemini.

Prompt example for Gemini inference with 1 shot RAG using cosine similarity

Input:

Your task is to answer the physics questions. The mathematical formulas are provided in Latex code. There are some related questions and their answers you may find helpful.\nHere are the examples:\n**Question**: In his famous work on magnetic monopoles Paul Dirac has argued that if just one magnetic monopole existed in the Universe, all electric charges should be multiple of a specific elementary electric charge, whose magnitude is related to the magnetic charge of that monopole. Historically, it is the first hypothesis in physics, which explains the existence of the elementary electric charge. Consider the system described in Question 6, assuming that all magnetic monopoles existing in the Nature have magnetic charges of the same magnitude, $+q_m$ and $-q_m$ respectively. By applying the concepts of quantum physics to the motion of electrically charged particle along the circular orbit, derive a relationship between the elementary electric charge e, assumed to be the charge of the electron, and the magnetic charge q_m of the monopole. Calculate q_m numerically.\n**Reference answer**: According to the Bohr's quantization rule, the change of the angular momentum along a specified axis is a multiple of the reduced Plank's constant: $ \\Delta L_z=n \\hbar=n \\frac{h}{2 \\pi} $ By utilizing the expression for $\\Delta L_z$, derived in the last question, we obtain: $ \\frac{\\mu_0 q_m q_e}{2 \\pi}=\\frac{n h}{2 \\pi} \\quad \\text { or } \\quad q_e=n \\frac{h}{\\mu_0 q_m} . $ This result implies that the electric charge is multiple of a universal elementary electric charge $\\frac{h}{\\mu_0 q_m}$. If we identify the elementary electric charge with the charge of the electron, we can find the magnitude of the magnetic charge of the monopole: $ q_m=\\frac{h}{\\mu_0 e}=3.29 \\times 10^{-9} \\mathrm{~A} \\cdot \\mathrm{m} $\n\n\n**The question that you need to solve is**:\nWave-particle duality, which states that each particle can be described as a wave and vice versa, is one of the central concepts of quantum mechanics. In this problem, we will rely on this notion and just a few other basic assumptions to explore a selection of quantum phenomena covering the two distinct types of particles of the microworld-fermions and bosons.Consider a particle of mass m moving in a one-dimensional potential well, where its potential energy $V(x)$ is given by $V(x)= \\begin{cases}0, & 0 \\leq x \\leq L \\ \\infty, & x<0 \\text { or } x>L\\end{cases}$.While classical particle can move in such a potential having any kinetic energy, for quantum particle only some specific positive discrete energy levels are allowed. In any such allowed state, the particle can be described as a standing de Broglie wave with nodes at the walls.Determine the minimal possible energy $E_{\\min }$ of the quantum particle in the well. Express your answer in terms of m, L, and the Planck's constant h.\n\nResponse with the FINAL answer to the question to get a higher score as possible as you can, rather than only give directions or suggestions for solving the problem. Do NOT use the conditions in the example questions to solve the question.

Output:

Assistant: The minimal possible energy of the quantum particle in the well corresponds to the ground state, which is the lowest energy level. \n\nHere's how we can determine it:\n\n1. **De Broglie Wavelength:** The particle can be described as a standing de Broglie wave with nodes at the walls. This means the length of the well, *L*, must accommodate half a wavelength of the wave. Therefore, the wavelength of the particle in the ground state is: $\lambda = 2L$\n\n2. **De Broglie Relation:** The De Broglie relation connects the momentum (*p*) of a particle to its wavelength (*λ*): $p = h/\lambda$\n\n3. **Kinetic Energy:** The kinetic energy of the particle is related to its momentum: $KE = p^2/2m$\n\n4. **Total Energy:** Since the potential energy is zero inside the well, the total energy *E* is equal to the kinetic energy: $E = KE = p^2/2m$\n\n5. **Substituting:** Now we can substitute the De Broglie relation (2) and the wavelength (1) into the energy equation (4): $E = (h^2/\lambda^2)/(2m) = (h^2/(2L)^2)/(2m)$\n\nTherefore, the minimal possible energy of the quantum particle in the well is:\n\n**$E_{min} = h^2/(8mL^2)$** \n

Figure 12: An example for Gemini inference with 1-shot RAG using cosine similarity.

E Multimodality Example

Questions: In a simple model, an atomic nucleus can be thought of as a ball consisting of closely packed nucleons [see Figure 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 Figure 1(b)]. (0.3 points)

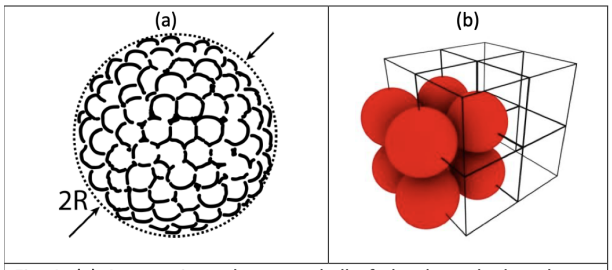

Fig. 1. (a) An atomic nucleus as a ball of closely packed nucleons. (b) The SC packing.

Figure 13: The Image from the Question.

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 $f = \frac{V_N}{a^3} = \frac{\pi}{6} \approx 0.52$

GPT-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.

F Runtime Analysis

Algorithm	Pure Retrieval (s)	Retrieval + Embedding (s)
Contriver	23	255
Dragon+	33	155
Cosine similarity	126	275
BM25	7	7.4

Table 8: Performance Comparison of Different RAG Algorithms in Second.

G Examples of answer errors

G.1 Error Type 1: Give guidelines instead of direct answer

Problem A ball is launched at speed v at angle θ above the horizontal toward a vertical wall a distance$L$ away. It bounces elastically off the wall and falls back to its launch point. What was its initialspeed?

Reference Answer $\frac{\sqrt{2Lg}}{\sqrt{\sin(2\theta)}}$

GPT-3.5's response To solve this problem, we can consider the motion in two parts: the motion from the launch point to the wall, and then the motion from the wall back to the launch point.For the first part of the motion (from the launch point to the wall), we can consider the horizontal motion and the vertical motion separately. In the horizontal direction, the ball travels a distance L with a constant horizontal velocity. In the vertical direction, the ball travels a vertical distance h, given by: $h = \frac{1}{2}gt^2$ where t is the time taken to reach the wall, and g is the acceleration due to gravity.Since the ball lands back at the launch point, the total time taken for the entire motion is $2t$. The vertical speed of the ball when it reaches the wall can be calculated as follows: $v_y = gt$. From the kinematic equation, we know that the vertical component of velocity when hitting the wall is equal to the initial vertical component of velocity multiplied by -1 (due to reflection of velocity), thus: $v\sin(\theta) = -gt$. Solving for t gives: $t = \frac{v\sin(\theta)}{g}$.Since an elastic collision preserves kinetic energy, the speed of the ball upon reflection is the same as the incoming speed vector. Therefore, the total horizontal distance traveled by the ball is $2L$. The total horizontal distance can be calculated as follows: $2L = v\cos(\theta) \cdot 2t$. Substitute the expression for t: $2L = v\cos(\theta) \cdot \frac{2v\sin(\theta)}{g}$. Solving for v gives: $v = \sqrt{\frac{gL}{\sin(2\theta)}}$.Therefore, the initial speed of the ball is $\sqrt{\frac{gL}{\sin(2\theta)}}$. **Score: 4**

Example 1 reterieved by cosine similarity A projectile is launched across flat ground at an angle θ to the horizontal and travels in the absence of air resistance. It rises to a maximum height H and lands a horizontal distance R away. What is the ratio H/R ? Reference answer: $\frac{1}{4}\tan\theta$.

Example 2 reterieved by cosine similarity A cannonball is launched with initial velocity of magnitude v_0 over a horizontal surface. At what minimum angle $\theta_{\min}$ above the horizontal should the cannonball be launched so that it rises to a height H which is larger than the horizontal distance R that it will travel when it returns to the ground? Reference answer: $\theta_{\min} = 76^\circ$.

GPT-3.5's response using two-shot Retrieval-Augmented Generation We can break down the motion of the ball into horizontal and vertical components. When the ball hits the wall and bounces back, the vertical component of its velocity remains the same while the horizontal component changes direction. Let v_x be the horizontal component of the initial velocity, and v_y be the vertical component of the initial velocity. Since the motion is symmetrical, the time taken to return to the launch point is twice the time taken to reach the wall:$t_{\text{total}} = 2 \cdot t_{\text{to wall}}$The time taken to reach the wall can be found using the horizontal distance L and the x-component of the initial velocity:$L = v_x \cdot t_{\text{to wall}}$From this, we solve for $t_{\text{to wall}}$:$t_{\text{to wall}} = \frac{L}{v_x}$Now, considering the vertical motion, at the maximum height H the vertical component of the velocity is zero, hence:$v_y^2 - 2 \cdot g \cdot H = 0$where g is the acceleration due to gravity. As the initial velocity is the magnitude of the total velocity:$v = \sqrt{v_x^2 + v_y^2}$From the above two equations we can solve for v. **Score: 3**

Analysis In this case, the LLM only provided advisory comments, stating 'From the above two equations we can solve for v'. without continuing to answer the question, resulting in a loss of some points.

G.2 Error Type 2: Misguided by the retrieved examples.

Problem The change in the gas temperature due to the adiabatic expansion and contraction, as a result of the sound wave, takes the form:$T(x,t) = T_0 - T_1(x)\cos(\omega t)$.Compute the amplitude $T_1(x)$ of the temperature oscillations in terms of T_0, γ, a, k and x.

Reference Answer Given: - Temperature perturbation due to the sound wave: $T(x,t) = T_0 - T_1(x)\cos(\omega t)$, - Adiabatic constant: γ, - Displacement amplitude: a, - Wave parameters: k and ω. We know that for adiabatic processes, $pV^\gamma = \text{const}$, and for ideal gases, $pV = nRT$ where n is the number of moles and R is the gas constant. From the ideal gas law, we have $p = \frac{nRT}{V}$. Let's consider the equilibrium state where $p = p_0$, $V = V_0$, and $T = T_0$. Assuming small temperature perturbation from the equilibrium state: $T = T_0 + T_1(x)\cos(\omega t)$. Substitute into the ideal gas law: $\frac{nR(T_0+T_1(x)\cos(\omega t))}{V_0} = p_0$. Expanding this expression to first order in the perturbations, we get: $T_1(x) = -\gamma T_0 a \sin(kx)$. Therefore, the amplitude $T_1(x)$ of the temperature oscillations is $-\gamma T_0 a \sin(kx)$ in terms of T_0, γ, a, k, and x.

GPT-3.5's response (Score: 6) In a thermoacoustic system where the gas oscillates due to a standing sound wave, and there is weak thermal interaction between the gas and the tube, let us analyze the temperature changes at specific points on the tube. At point A, located at the edge of the tube, the gas parcel experiences maximum compression and minimum volume. As the gas undergoes an adiabatic expansion, it cools down, leading to a decrease in the tube's temperature at this point over time. At point B, the central part of the tube, the gas parcel is at its equilibrium position, neither at maximum compression nor expansion. Here, the temperature tends to remain constant as the gas parcel neither gains nor loses significant heat to the tube. At point C, another edge of the tube, the gas parcel is at maximum expansion and volume. Upon adiabatic compression, the gas heats up, resulting in an increase in the tube's temperature at this point. Therefore, in this scenario of weak thermal interaction, over a prolonged period, the temperature at point A decreases, remains unchanged at point B, and increases at point C.

Example 1 retrieved by cosine similarity Consider a longitudinal wave traveling along the x direction in the rod. For a cross section at x when the rod is unstressed, let $\xi(x,t)$ be its displacement at time t and assume $\xi(x,t) = \xi_0 \sin k(x - ut)$ where ξ_0 and k are constants. Determine the corresponding velocity $v(x,t)$, strain $S(x,t)$, and pressure $p(x,t)$ as a function of x and t. Reference answer: The velocity v is related to the displacement ξ as in a simple harmonic motion (or a uniform circular motion, as shown in Figure A2) of angular frequency $\omega = ku$. Therefore, if $\xi(x,t) = \xi_0 \sin k(x - ut)$, then $v(x,t) = -ku\xi_0 \cos k(x - ut)$. The strain and pressure are related to velocity as in Problem (a). Hence, $S(x,t) = -v(x,t)/u = k\xi_0 \cos k(x - ut) p(x,t) = \rho u v(x,t) = -k\rho u^2 \xi_0 \cos k(x - ut) = -YS(x,t) = -kY\xi_0 \cos k(x - ut)$.

Example 2 retrieved by cosine similarity In the uniform background magnetic field, theplasma density is nonuniform in x. For simplicity, we assume that the temperature and the distribution of the ions and electrons are the same. Thus, the plasma pressure can be expressed as $p(x) = kT\left[n_i(x) + n_e(x)\right] = 2kTn(x) = 2kT\left(n_0 + \alpha x\right)$, Where B, T, k, n_0, and α are positive constants, $n_i(x)$ and $n_e(x)$ are the number densities of the ions and electrons. The diamagnetic field is given by $B_{mx} = \mu_o M$. Please explain the generation mechanism of the electric current by a schematic drawing. Reference answer: Since the magnetic field and the plasma are uniform z, the orbits of ions and electrons can project into in the $x - y$ plane. From the results of sub_question_number: 1, we know that an ion has a left-hand circular motion and an electron has a right-hand circular motion. Due to the linear increase of the plasma density in x, the number of ions with upward motion is less than that with downward motion at a given x position, which leads a net upward ion flow. Similarly, electrons have a net downward flow. Combining the ion and election flows, we have a net upward electric current as illustrated below in schematic drawing.

GPT-3.5's response using two-shot Retrieval-Augmented Generation (Score: 4) Given: - Temperature perturbation due to the sound wave: $T(x,t) = T_0 - T_1(x)\cos(\omega t)$, - Adiabatic constant: γ, - Displacement amplitude: a, - Wave parameters: k and ω. We know that for adiabatic processes, $pV^\gamma = \text{const}$, and for ideal gases, $pV = nRT$ where n is the number of moles and R is the gas constant. From the ideal gas law, we have $p = \frac{nRT}{V}$. Let's consider the equilibrium state where $p = p_0$, $V = V_0$, and

$T = T_0$. Assuming small temperature perturbation from the equilibrium state: $T = T_0 + T_1(x)\cos(\omega t)$. Substitute into the ideal gas law: $\frac{nR(T_0+T_1(x)\cos(\omega t))}{V_0} = p_0$. Expanding this expression to first order in the perturbations, we get: $T_1(x) = -\gamma T_0 a \sin(kx)$. Therefore, the amplitude $T_1(x)$ of the temperature oscillations is $-\gamma T_0 a \sin(kx)$ in terms of T_0, γ, a, k, and x.

Analysis The original question was relatively short, but the retrieved questions are quite lengthy, inevitably focusing more attention on the retrieved content. After the RAG process, it is evident that the LLM listed conditions that did not exist in the original question.

G.3 Error Type 3: Wrong Retrieval.

Problem his part deals with the usage of microwave radiation (radiated by the magnetron antenna into the food chamber) for cooking, that is, heating up a lossy dielectric material such as water, either pure or salty (which is our model for, say, soup). An electric dipole is a configuration of two equal and opposite electric charges q and $-q$ a small distance d apart. The electric dipole vector points from the negative to the positive charge, and its magnitude is $p = qd$. A time-dependent electric field $\vec{E}(t) = E(t)\hat{x}$ is applied on a single dipole of moment $\vec{p}(t)$ with constant magnitude $p_0 = |\vec{p}(t)|$. The angle between the dipole and the electric field is $\theta(t)$.Write expressions for both the magnitude of the torque $\tau(t)$ applied by the electric field on the dipole and the power $H_i(t)$ delivered by the field to the dipole, in terms of $p_0, E(t), \theta(t)$ and their derivatives.

Reference Answer The torque at time t is given by $\tau(t) = -qd\sin[\theta(t)]E(t) = -p_0\sin[\theta(t)]E(t)$,hence the instantaneous power delivered to the dipole by the electric field is $H_i(t) = \tau(t)\dot{\theta}(t) = -p_0E(t)\sin\theta(t)\dot{\theta}(t) = E(t)\frac{d}{dt}(p_0\cos\theta(t)) = E(t)\frac{dp_x(t)}{dt}$.

GPT-3.5's response (Score: 9) The torque $\tau(t)$ applied by the electric field on the dipole can be expressed as the cross product of the electric dipole moment $\vec{p}(t)$ and the electric field $\vec{E}(t)$:

$$\boldsymbol{\tau}(t) = \vec{p}(t) \times \vec{E}(t)$$

Expanding the cross product using vector algebra gives:

$$\boldsymbol{\tau}(t) = p_0\hat{p}(t) \times E(t)\hat{x} = p_0E(t)\sin(\theta(t))\hat{z}$$

where $\hat{p}(t)$ is the unit vector pointing along the direction of the dipole moment $\vec{p}(t)$ and $\hat{z}$ is the unit vector along the z-axis.The power $P(t)$ delivered by the field to the dipole can be calculated using the dot product of the torque $\boldsymbol{\tau}(t)$ and the angular velocity $\boldsymbol{\omega}(t)$ of the dipole:

$$P(t) = \vec{\tau}(t) \cdot \vec{\omega}(t)$$

As the electric field $\vec{E}(t)$ is time-dependent and not directly causing the dipole to rotate, the power delivered to the dipole is due to changes in the potential energy of the system, which can be related to the time derivative of the dipole moment $\vec{p}(t)$. Therefore, the power $H_i(t)$ delivered by the field to the dipole can be expressed as:

$$H_i(t) = \frac{d}{dt}(\vec{p}(t) \cdot \vec{E}(t))$$

By substituting the given values, we obtain the expression for the power:

$$H_i(t) = \frac{d}{dt}(p_0E(t)\cos(\theta(t)))$$

Example 1 retrieved by cosine similarity In the figure below, a simplified scheme of a mass-spectrometer is given. It is a device for measuring the masses of molecules. The substance under investigation is ionised by heating up to a temperature T on a hot filament (molecules undergo a single-electron ionisation). The ions are accelerated using voltage U. At first, let us neglect the thermal energy of the ions ($eU \gg kT$, where e is elementary charge and k - the Boltzmann's constant). A narrow beam of accelerated ions enters a region with magnetic field. For the sake of simplicity, let us assume that the region has a rectangular shape, and the magnetic field is homogeneous inside it. The magnetic fields deflects the ions and depending on their mass, they may hit the detector. Let us assume that those ions which hit the centre of the detector enter and exit the region with magnetic field perpendicularly with its boundary, and the distance between the entry and exit points is l (see figure). 1) Express the mass M of those ions which hit the centre of the detector via the quantities $B\,l, U$ and e.Reference answer:The trajectory of a charged particle in the magnetic field is circle of radius $R = l/\sqrt{2}$. Lorenz force is responsible for the acceleration, Bev $= Mv^2/R$ hence $BeR = p$. Substituting $p^2 = 2MUe = B^2e^2R^2$, we obtain $M = B^2l^2e/4U$.

Example 2 retrieved by cosine similarity One way of measuring the atmospheric electric field, and hence σ_0, is with the system shown in the diagram. A pair of metal quadrants, insulated from ground but connected to each other, are mounted just underneath a grounded uniformly rotating disk with two quadrant-shaped holes cut in it. (In the diagram, the spacing has been exaggerated in order to show the arrangement.) Twice in each revolution the insulated quadrants are completely exposed to the field, and then (1/4 of a period later) are completely shielded from it. Let T be the period of revolution, and let the inner and outer radii of the insulated quadrants be r_1 and r_2 as shown. Take $t = 0$ to be an instant when the insulated quadrants are completely shielded. Obtain expressions that give the total charge $q(t)$ induced on the upper surface of the insulated quadrants as a function of time between $t = 0$ and $t = T/2$, and sketch a graph of this variation. [The effects of the atmospheric ion current can be ignored in this situation.]

GPT-3.5's response using two-shot Retrieval-Augmented Generation (Score: 4) The torque applied by an electric field on a dipole is given by $\boldsymbol{\tau}(t) = \vec{p}(t) \times \vec{E}(t)$. Since $\vec{p}(t) = p_0 \cos(\theta(t))\hat{y}$ and $\vec{E}(t) = E(t)\hat{x}$, the torque would be:

$$\boldsymbol{\tau}(t) = p_0 E(t) \cos(\theta(t))\hat{z}$$

The power P delivered by the electric field to the dipole is given by $\frac{dW}{dt} = \vec{F} \cdot \vec{v}$, where $\vec{F}$ is the force on the dipole and $\vec{v}$ is its velocity. Since power is force times velocity, we get:

$$P = \vec{F} \cdot \vec{v} = \boldsymbol{\tau} \cdot \vec{\omega}$$

Where $\vec{\omega}$ is the angular velocity of the dipole. Substituting $\boldsymbol{\tau}(t) = p_0 E(t) \cos(\theta(t))\hat{z}$ and $\vec{\omega}(t) = -\dot{\theta}(t)\hat{z}$, we find:

$$H_i(t) = \boldsymbol{\tau}(t) \cdot \vec{\omega}(t) = -p_0 E(t)\dot{\theta}(t) \cos(\theta(t))$$

.

Analysis Although the original topic is about wave radiation, and the searched topics are also related to waves, the specific content and background of the topics are different: Example 1 describes a simplified model of a mass spectrometer. Example 2 discusses a method for measuring the electric field of the atmosphere.

H Additional Results

H.1 Experimental results of problems with different difficulty levels.

We provide the number of questions of different difficulty levels, as shown in Table 9. The number of questions whose answer scores are improved from zero to non-zero after RAG are shown in Table 11. Average score of LLMs with RAG over different difficulty level problems are provided in Table 10.

Category	Easy	Medium	Hard	Superhard
Count	51	74	148	117

Table 9: Distribution of Question Difficulty.

Model	Method	Easy	Medium	Hard	Superhard
GPT-3.5	Base	9.41	5.89	3.76	1.15
GPT-3.5	BM25	6.51	4.74	3.55	2.45
GPT-3.5	Dragon+	6.63	4.08	3.68	2.38
GPT-3.5	Contriver	6.35	4.85	3.72	2.48
GPT-3.5	Cos	6.57	5.04	3.75	2.62
GPT-4	Base	8.45	7.15	6.14	4.91
GPT-4	BM25	8.51	6.81	5.78	4.42
GPT-4	Dragon+	8.53	7.07	5.60	4.38
GPT-4	Contriver	8.94	6.70	5.29	4.60
GPT-4	Cos	8.45	6.45	5.41	4.17

Table 10: Average score of GPT-3.5 and GPT4 with RAG in different difficulty level problems.

Model	BM25	Cos	Dragon+	Contriver	Average
GPT-4	6	7	5	9	6.75
GPT-3.5	18	18	24	19	19.75
Gemini	26	22	23	22	23.25
DeepSeek-Math	10	11	12	12	11.25

Table 11: Number of questions whose answer scores are raised from zero to non-zero after RAG.