Chain-of-Thought Physics Benchmark for Reasoning Models

Filomela V. Gerou[a,b*] **Nesar S. Ramachandra**[a] **Azton Wells**[a]

[a]Argonne National Laboratory, 9700 S Cass Ave, Lemont, IL 60439, USA
[b]University of Michigan, Ann Arbor, MI 48109, USA

Abstract

We introduce a comprehensive benchmarking framework designed to rigorously assess large language model (LLM) performance on structured reasoning tasks in advanced physics domains. The benchmark consists of chain-of-thought multiple-choice (MCQ) problems derived from advanced mathematical and physics texts and primary research literature. Central to this framework is a novel automated question-generation pipeline, which constructs problems from curated analytical sources using a predefined rubric, followed by expert-informed filtering to ensure conceptual rigor and relevance. Model evaluation is conducted via a multi-layered assessment protocol incorporating a reward–penalty grading scheme, pedagogical reasoning feedback, and a jury of large language models (LLM-jury) evaluating both logical coherence and correctness. We enable fine-grained analysis of reasoning pathways in multi-step physics problem solving. By explicitly tracking intermediate reasoning steps, the framework advances interpretability in physics-relevant AI systems. We observe reasoning improvement across three consecutive generations of OpenAI, Anthropic, and Google LLMs, with all normalized scores remaining below 0.4. This feedback-driven, multi-step, reasoning-focused benchmark in physics closely resembles real-world human-LLM interaction.

1 Introduction

Existing benchmarks that assess mathematical problem solving and reasoning ability such as GPQA [Rein et al., 2023], GSM8K [Cobbe et al., 2021], MATH [Hendrycks et al., 2021], and AIME [Shi et al., 2025] have demonstrated that LLMs can perform well on general knowledge and competition-style mathematics problems. Mathematical benchmarks like FrontierMath [Glazer and et al., 2024] introduce well-posed problems but lack the explicit multi-step reasoning processes that are indispensable for physics and astronomy problem solving. These consist of automated benchmarks of low difficulty scores or unscalable complex problem sets. Astronomy-focused benchmarks, such as AstroBench [Ting et al., 2025], have emphasized factual recall and conceptual problems. Similarly, large-scale physics benchmarks, including UGPhysics [Xu et al., 2025] and GPQA [Rein et al., 2023], are primarily constructed from university coursework. While such datasets provide broad coverage, they are publicly available, raising concerns of training-data contamination [Xu et al., 2025], and they predominantly test knowledge application as opposed to creative or generative problem-solving ability. Reasoning-oriented resources such as Scheherazade [Miner et al., 2025] present isolated questions with limited continuity, thereby restricting evaluation of sustained inference. Evaluation approaches within MCQ benchmarks, such as AstroBench [Ting et al., 2025], Humanity's Last Exam [Center for AI Safety et al., 2026], and ARC-AGI [Chollet et al., 2026], collapse reasoning into a single accuracy-focused outcome metric. Simultaneously, rubric-based assessments, such as

*fgerou@umich.edu

Conference on Physics and AI at Stanford University (PAI 2026).

UGphysics [Xu et al., 2025] provide structured, but narrow criteria. IMO benchmarks [Luong et al., 2025] set a high bar for mathematical difficulty, but are not scalable. Aforementioned benchmarks evaluate LLMs on isolated questions. This does not reflect real-world use cases, where users provide feedback, hints, and expressions of uncertainty in a closed-loop interaction with the system.

These limitations underscore the scarcity of domain-specific benchmarks that capture the complexity of problem-solving reasoning in physics and astronomy. To address this gap, we present a scalable benchmark derived directly from research literature, designed to evaluate not only correctness, but also efficiency, iterative reasoning, integration of knowledge, and the capacity for refinement and reformulation in scientific problem solving.

2 Dataset Formulation

Our question generation pipeline produces a dataset of chained multiple-choice questions demanding step-by-step problem solving. To achieve both scalability and rigor, the pipeline is divided into three stages: (i) source literature retrieval by human experts, (ii) automated generation of candidate questions, and (iii) feedback- and reward-penalty-driven refinement guided by an LLM-question-judge. We generate a dataset of 160 problem chains, corresponding to 800 individual questions. A chain represents a complete, multi-step problem and consists of 5 sub-questions, each capturing a distinct stage of the reasoning process required in solving the problem. Models are evaluated at every step, and progression through a chain requires all prior steps to be answered correctly. Our dataset is comprised of 65% application and 35% conceptual questions. Conceptual questions test whether a model can correctly identify the relevant ideas or laws for a physics problem, while application questions assess the ability to execute those concepts, laws, or methods to obtain the correct result. Figure 1 presents an example dataset question.

In the few-shot stage, we manually construct a seed dataset of high-complexity questions ranging from graduate to frontier research difficulty, alongside a curated corpus of 30 papers and textbooks. Grounding question generation in primary literature reduces trivial substitution and direct knowledge-retrieval items, enforcing reasoning depth commensurate with active research practice. Following Center for AI Safety et al. [2026], we prioritize problems demanding expert-level understanding at both foundational and frontier levels. Questions are grounded in source papers where answers are non-trivial yet traceable, reducing the risk of ground-truth invalidity identified by Skarlinski et al. [Skarlinski et al., 2025]. In the automated stage, each candidate question is evaluated by the LLM-question-judge against a detailed rubric. This LLM-question-judge provides a diagnostic score and targeted feedback that drives a reward–penalty refinement loop. This ensures questions require logically coherent, multi-step reasoning rather than formulaic application. Questions exceeding a quality threshold of 0.8 are admitted, with roughly half of the generated questions rejected at first attempt. This threshold accounts for LLM-Judge scoring variability. We, thus, increase the likelihood of generating a dataset exhibiting progressive reasoning with rigorously justified ground-truth solutions.

3 Model Evaluation

We develop a novel assessment methodology combining deterministic, penalty-based scoring with adjudication by a jury of LLM-as-evaluator-judges, while tracking response latency and token usage to capture efficiency alongside accuracy. Answer score deductions rely on the number of retries required, while transitions into subsequent evaluation phases are not penalized, as they may reflect valid but unlisted solution strategies. Reasoning scores are penalized for the inclusion of irrelevant rationales or inconsistencies between the proposed rationale and the selected answer. All rationales are stored to enable tracking of reasoning progression and investigation of correlations between retry count and thematic coherence. This provides insight into the fundamental physics understanding and problem decomposition capabilities of the test models. Taken together, these metrics reduce dependence on the model's stated rationale as a proxy for its internal reasoning process [Radhakrishnan et al., 2023]. Evaluation proceeds through two sequential feedback stages.

In the first stage, the tested model generates a rationale for each sub-question in the chain. This rationale is compared against ground-truth rationales using the SequenceMatcher similarity score, with a threshold of 0.7 (Figure 1) allowing small deviations from descriptions of physical laws or mathematical methods listed in the dataset's ground-truths. The dataset includes predetermined

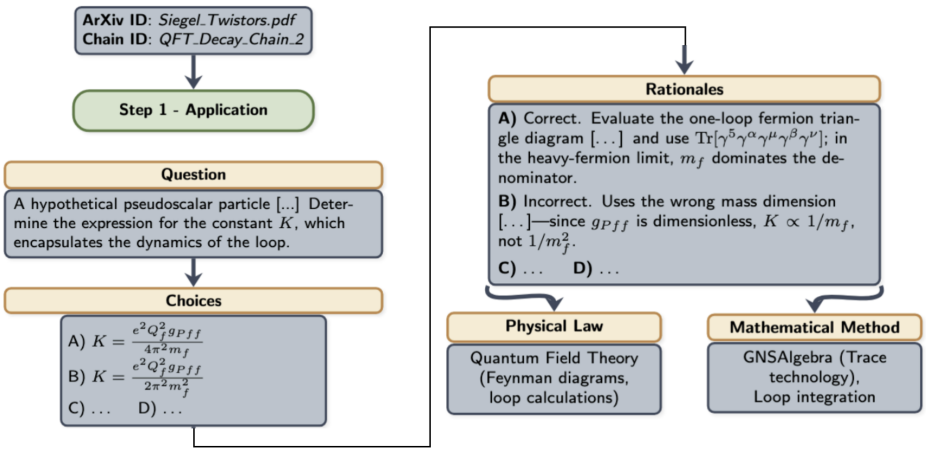

Figure 1: The structure of a truncated example problem chain; the first step (application-type) shows answer choices with their corresponding rationales, and relevant physical laws and methods.

plausible error sources corresponding to each answer choice. Ground-truth rationales encode the reasoning required for correct problem solving, while distractor rationales represent plausible but erroneous reasoning pathways. If no match is achieved, answer selections are not scored and the model is automatically prompted to reformulate its reasoning. This stage enables targeted feedback without LLM-assisted evaluation, thereby reducing computational overhead. Each model begins with the maximum score on two axes, reflecting final answer selection and rationale quality. The answer score for each chain is calculated as $0.2\left(1 - \sum_{r=0}^{2}\left(0.1r^{3/2} + 0.1r\right)\right) - 0.01l$, where r denotes the retry count and l the latency. For each sub-question, the maximum score of 1 is reduced by a nonlinear penalty term $r^{3/2}$ chosen to impose a larger penalty on secondary retries. Scores may take negative values, primarily due to the contribution of un-normalized latency. The 0.2 factor normalizes the aggregate score from the five sub-questions into a single chain-level score. Invalid responses, including incomplete or erroneous rationales accompanied by a correct selected answer, are flagged. This mitigates the risk of inflated scores from elimination, selection bias, or guesswork [Pezeshkpour and Hruschka [2023]]. The larger number of scoring permutations compared with a conventional multiple-choice benchmark, reduces the minimum achievable score from 0.25 to 0.04. This is due to the accumulative scoring from multiple sub-questions and reformulations. Consequently, the benchmark's feedback-loop-driven format is incompatible with raw answer scoring.

In the second stage, upon exhaustion of hard-coded retries, evaluation is escalated to a LLM-jury: Google Gemini-2.5-pro, OpenAI GPT-o3, and Anthropic Claude-sonnet-4 to ensure broad coverage across all tested model families, minimizing systematic adjudication bias [Verga et al., 2024]. Self-adjudication bias is limited, as each assessment is an independent inference pass, with no retained memory. Each judge applies explicit criteria to assess reasoning soundness, correctness of method application, and appropriate use of mathematical tools. This includes independently attempting to solve the sub-question using the outlined reasoning methodology in the tested model's response, penalizing rationale falsehood [Chen et al., 2025]. This stage additionally enables identification and documentation of novel, but valid solution strategies. A summary of the evaluator judge's responses in the form of collective pedagogical feedback is returned to the model to prompt rationale reformulation. This stage emulates the student-educator interaction of the physics education process. Here, we provide the model with the opportunity to self-assess and improve its approach when engaging with scientific research. The utility of step-wise answer formation and iterative feedback for model reasoning is explored in Uesato et al. [2022], Zhou and Zhang [2025]. If, after, a heuristically chosen, $N = 3$ number of feedback iterations, the response fails to pass both string comparison and jury assessment, the model receives a score of zero for remaining sub-questions in the chain.

[1]We evaluate 8 LLMs from 3 frontier model families spanning 3 generations (Figure 2). We collect 15 independent responses per chain, amounting to 2250 total inputs per model. This reduces estimator variance.

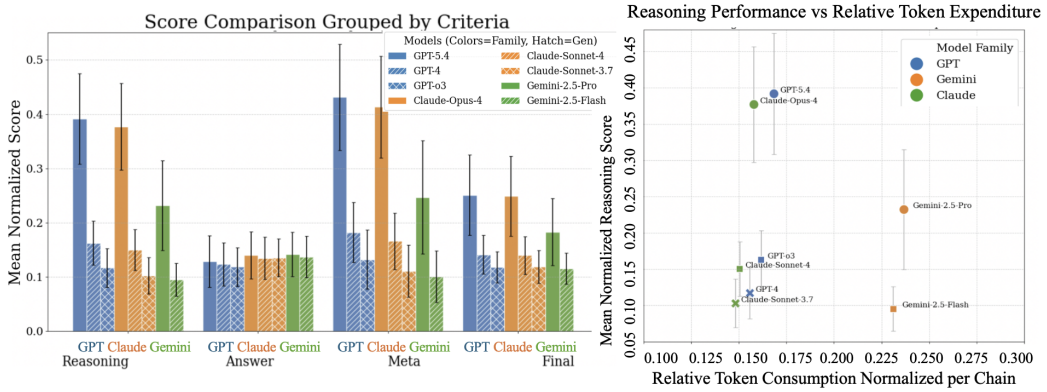

Figure 2: Left: total model performance across reasoning (rationale quality), answer accuracy, meta (retries, latency, token usage), and final (average of answer and reasoning scores). Right: reasoning score vs. token usage, normalized per chain across models and shots.

4 Results

All evaluation metrics are min-max normalized within 0-1 with a consistent framework across models and shots prior to aggregation. In Figure 2, we observe an average $59\% \pm 13.5\%$ decrease in model reasoning performance from GPT-5.4, Claude-opus-4, Gemini-2.5-pro to GPT-4, Claude-sonnet-4, and Gemini-2.5-flash, respectively. The highest reasoning score is achieved by GPT-5.4 at 0.39 ± 0.08, followed by Claude-opus-4 at 0.37 ± 0.08 and Gemini-2.5-pro at 0.23 ± 0.08. Answer accuracy follows a similar generation-wide trend, with Claude-opus-4 and Gemini-2.5-pro achieving the highest score at 0.14 ± 0.04 and GPT-5.4 following at 0.13 ± 0.05. Notably, the Gemini models achieve comparatively lower scores, while exhibiting higher token use [Team, 2024]. OpenAI and Anthropic models exhibit relatively consistent token expenditure while showing improved performance across generations. Topic-wise analysis of normalized reasoning scores indicates relatively uniform performance across physics domains, with variations constrained to within ± 0.05. Performance is highest in astrophysics and lowest in quantum mechanics. Empirical analysis of rationale-tracking logs shows that models often reproduce erroneous rationales with little or no modification during retries or revert to their original incorrect reasoning if a revision is rejected. Additionally, we identify instances in which the model articulates a methodologically correct rationale with appropriate input values, yet produces inconsistent numerical results across retries, suggesting deficiencies in algebraic execution [Song et al., 2026]. Simultaneously, correct final answers without the corresponding mathematical approaches may be stated, which may indicate inconsistency between internal computation and reported rationale [Radhakrishnan et al., 2023].

5 Conclusions

Our contributions include: (1) a domain-specific reasoning benchmarks for physics/astronomy research; (2) a scalable, feedback-driven dataset generation process; (3) a hybrid evaluation combining semantic similarity with pedagogical jury-of-judges feedback; and (4) increased LLM interpretability through a physics-relevant assessment. Our work moves toward increased transparency in understanding a model's ability to reproduce underlying physics in research settings. Planned extensions of this work include evaluating autonomous AI agents in physically grounded problem-solving contexts and utilizing our feedback loops for a reinforcement learning platform. While our framework mitigates shortcomings of prior benchmarks, the absence of systematic human expert evaluation and the computational costs of LLM-based evaluation present limiting factors. Further, while MCQs rarely reflect real-world use cases, they remain a practical validation route when divergent reasoning paths make open-ended evaluation intractable. Despite these challenges, this framework establishes a modular and extensible path for evaluating scientific reasoning as an iterative process rather than a final-answer benchmark.

Acknowledgments

This material is based upon work supported by Laboratory Directed Research and Development (LDRD) funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DEAC02-06CH11357. We gratefully acknowledge the use of Argo, an LLM serving framework developed to support research and operations at Argonne National Laboratory. By accepting the article for publication, the publisher acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The repository with the relevant codes for the benchmark and the comprehensive dataset is available at https://github.com/filomelavgerou/BeyondFinalAnswers.

References

Center for AI Safety, Scale AI, and HLE Contributors Consortium. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649:1139–1146, 2026. doi: 10.1038/s41586-025-09962-4. URL https://arxiv.org/abs/2501.14249.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think, 2025. URL https://arxiv.org/abs/2505.05410.

François Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. Arc prize 2025: Technical report, 2026. URL https://arxiv.org/abs/2601.10904.

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

Elliot Glazer and et al. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint*, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf.

Thang Luong, Dawsen Hwang, Hoang H. Nguyen, Golnaz Ghiasi, Yuri Chervonyi, Insuk Seo, Junsu Kim, Garrett Bingham, Jonathan Lee, Swaroop Mishra, Alex Zhai, Clara Huiyi Hu, Henryk Michalewski, Jimin Kim, Jeonghyun Ahn, Junhwi Bae, Xingyou Song, Trieu H. Trinh, Quoc V. Le, and Junehyuk Jung. Towards robust mathematical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://aclanthology.org/2025.emnlp-main.1794/.

Stephen Miner, Yoshiki Takashima, Simeng Han, Sam Kouteili, Ferhat Erata, Ruzica Piskac, and Scott J Shapiro. Scheherazade: Evaluating chain-of-thought math reasoning in llms with chain-of-problems, 2025. URL https://arxiv.org/abs/2410.00151.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions, 2023. URL https://arxiv.org/abs/2308.11483.

Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Sam McCandlish, Sheer El Showk, Tamera Lanham, Tim Maxwell, Venkatesa Chandrasekaran, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Question decomposition improves the faithfulness of model-generated reasoning, 2023. URL https://arxiv.org/abs/2307.11768.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Yuanzhe Richard Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint*, 2023.

Yexuan Shi, Mingyu Wang, Yunxiang Cao, Hongjie Lai, Junjian Lan, Xin Han, Yu Wang, Jie Geng, Zhenan Li, Zihao Xia, Xiang Chen, Chen Li, Jian Xu, Wenbo Duan, and Yuanshuo Zhu. Aime: Towards fully-autonomous multi-agent framework, 2025. URL https://arxiv.org/abs/2507.11988.

M. Skarlinski, J. Laurent, A. Bou, and A. White. About 30% of humanity's last exam chemistry/biology answers are likely wrong. FutureHouse, 2025. URL https://www.futurehouse.org/research-announcements/hle-exam.

Peiyang Song, Pengrui Han, and Noah Goodman. Large language model reasoning failures, 2026. URL https://arxiv.org/abs/2602.06176.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Y.-S. Ting, T. D. Nguyen, T. Ghosal, R. Pan, H. Arora, Z. Sun, T. de Haan, N. Ramachandra, A. Wells, S. Madireddy, and A. Accomazzi. Astromlab 1: Who wins astronomy jeopardy!? *Astronomy and Computing*, 51:100893, 2025.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback, 2022. URL https://arxiv.org/abs/2211.14275.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models, 2024. URL https://arxiv.org/abs/2404.18796.

Xin Xu, Qiyun Xu, Tong Xiao, Tianhao Chen, Yuchen Yan, Jiaxin Zhang, Shizhe Diao, Can Yang, and Yang Wang. Ugphysics: A comprehensive benchmark for undergraduate physics reasoning with large language models, 2025. URL https://arxiv.org/abs/2502.00334.

Kuo Zhou and Lu Zhang. Step-wise formal verification for llm-based mathematical problem solving. *arXiv preprint*, 2025.