Benchmancy: Extending the Epistemic Validity of Scientific Reasoning Benchmarks

Anonymous ACL submission

Abstract

To mirror the increased interest in pursuing science as a source of rich, computational challenges for AI models, AI developers have begun developing new evaluations to assess the propensity of AI models to perform scientific reasoning tasks. However, these benchmarks are limited by their construction in their ability to represent the real-world parallels of scientific reasoning. This introduces an epistemic risk surface, stemming from construct validity issues inherent to these popular reasoning evaluations. We introduce benchmancy, a new evaluation framework designed to leverage existing benchmarks and enhance their representation of real-world tasks ex post facto. We instantiate benchmancy to enhance the epistemic validity of popular scientific reasoning benchmarks (MMLU, GPQA, Humanity's Last Exam, and FrontierScience) by expanding it through the inclusion of source attribution. We demonstrate that frontier LLMs are unable to accurately perform scientific reasoning tasks when prompted to provide evidence of its knowledge sourcing. We also show that benchmancy is able to reclaim saturated benchmarks by providing a more epistemically valid construct to evaluate.

1 Introduction

On May 14, 2026, the ubiquitous open-access research archive arXiv announced that it would impose a 1-year ban on authors who submit works that contain "incontrovertible evidence that the authors did not check the results of LLM generation." The listed examples of this type of oversight included plagiarized or biased content, meta-comments from the LLM, and incorrect references (Thomas G. Dietterich [@tdietterich], 2026). In spite of some localized pushback, this decision was widely applauded as a critical step towards defending the epistemic integrity of scholastic research. As the rate of AI-assisted publications rise significantly (Naddaf, 2026), so too does the number of reports exposing erroneous references. AI writing detection company GPTZero confirmed 100 hallucinated citations across 51 NeurIPS 2025 submissions (GPTZero, 2026) and found 50 submissions to ICLR 2026 (GPTZero, 2025) that contained confirmed hallucinated citations. This threat to trustworthy scientific knowledge is not limited to conference submissions, but also extends to the broader world of academic peer-review. One study conservatively estimated 0.1% (149,932) of citations in submissions to arXiv, bioRxiv, SSRN, and PubMed Central were hallucinated in 2025 alone (Zhao et al., 2026).

Barring directly connecting with the authors of any work, the citations present in peer-reviewed publications serve as the only publicly accessible vehicle for interpreting the chain of thought behind any scientific research effort. It is through citations that authors are able to expose what led them to their inspiration and carried them through the research process to produce an accurate and valid contribution to the scientific community at large. Trustworthy scientific research is only possible if all steps of the scientific process are verifiably measurable, documentable, and correct (Suri and Thais, 2025). Scientists are therefore dependent on the validity of published citations to justify trust in the knowledge produced by their peers (**?**). It is for these exact reasons that the generation of hallucinated or erroneous references poses a critical epistemic risk to the scientific community. Without verifiable source attribution, we risk losing a window into the minds of our peers, one that allows us to learn, replicate, and expand on their contributions.

While citations are curated and enumerated at the publication stage of the scientific research process, the necessity for transparent and accurate traceability does not stop here. As an epistemic venture, science is at its core expansive, expanding what was known yesterday to explain what is un-

known today. Every part of the scientific research process is grounded by some form of engagement with the compendium of scientific literature. Every part therefore requires its own guarantees of valid knowledge traceability. We pose that preventing hallucinated citations is insufficient for protecting the epistemic integrity of scientific research; what is necessary is ensuring valid source attribution. For this paper, we define source attribution as any mapping that ascribes intellectual content to its original source material. In this manner, source attribution extends beyond citation generation to the traceability underpinning queries of scientific reasoning. Thus, it is necessary to direct our attention to designing evaluations of AI systems that are fully representative of what constitutes scientific research.

The rise in interest by both scientists and AI researchers to develop AI systems to augment scientific research has naturally led to a growing desire to quantify how well these systems can produce valid scientific reasoning. In their release of FrontierScience titled "Evaluating AI's ability to perform scientific research tasks", OpenAI described reasoning as the "core of scientific work" (Wang et al.). Their intent behind the creation of FrontierScience was to develop a benchmark to "measure expert-level scientific capabilities" (Wang et al.). Here, we notice what is commonly known as a construct validity issue in which there exists a discrepancy between the evaluation aim (scientific reasoning) and the implemented task (Raji et al., 2021). By design, scientific reasoning benchmarks like FrontierScience only test AI models on their ability to work through self-contained, sometimes multi-step, theoretical questions. We argue that this construction is insufficient to properly evaluate scientific reasoning as the current form only captures one dimension of what constitutes real-world scientific reasoning, ignoring other critical aspects such as evaluations of source attribution. As humans and AI models are being pushed to operate in shared spaces and collaborate on vanguard research problems, it is imperative that evaluations that are both rigorous and realistic are developed. This means not only writing questions with the help of subject matter experts, but also designing tasks that are structurally representative of use cases "in the wild."

To correct the validity issues of contemporary scientific reasoning benchmarks, the ideal solution is to design new ones. However, the creation of a benchmark is a laborious endeavor and can often not be as impactful as intended. With the rapid rate of model development and improvement, benchmarks quickly tend towards states of stagnation or saturation. Even before saturation, it has been shown that a significant fraction of benchmarks are referenced by a limited number of different studies after its creation (Li et al., 2025). Additionally, building singular benchmarks often leads to further validity issues stemming from their over-specialization and consequently sparse coverage of the desired evaluation space. While contemporary benchmarks have known validity issues (Raji et al., 2021)(Salaudeen et al., 2025), they still remain valuable resources, the product of months of collaborative, careful query curation and harness design. As such, this paper introduces benchmancy: a cost-effective way to leverage and extend existing benchmarks to better measure real-world tasks. Instead of designing new benchmarks for scientific reasoning, we utilize the fact that scientific research tasks are often interconnected to apply a secondary harness to existing benchmarks ex post facto.

The contributions of the paper are as follows: **1)** We introduce benchmancy as a new, efficient means to extend the construct validity of existing benchmarks (Section 2) and contextualize its importance within the current benchmarking landscape (Section **??**). **2)** We construct a benchmancy harness for existing scientific reasoning benchmarks (MMLU (Hendrycks et al., 2021), GPQA (Rein et al., 2023), Humanity's Last Exam (Phan et al., 2025), and FrontierScience (Wang et al.)) to additionally measure source attribution (Section 3). **3)** We demonstrate the current limitations of large-language models to perform accurate source attribution for scientific reasoning tasks (Section 4). **4)** We show that benchmancy can revive saturated benchmarks and extend their use and construct validity.

2 Benchmancy

Benchmarks are an intuitive method to ascertain model aptitude on a desired task, measured against some predetermined standard, oftentimes a human baseline. Raji et al. (2021) defines benchmarks as "a particular combination of a dataset or sets of datasets (at least test data, sometimes also training data), and a metric, conceptualized as representing one or more specific tasks or sets of abilities, picked up by a community of researchers as

a shared framework for the comparison of methods." Within the AI community, benchmarks are then used as measurements of progress towards "general-purpose capabilities" or individual epistemic constructs (Raji et al., 2021). However, this assumption of valid extrapolation constitutes a construct validity issue. Benchmarks, as evaluations made up of independent queries, are therefore limited by construction in their ability to truly capture all facets of the evaluation target or construct (Salaudeen et al., 2025).

Literature critical of benchmarks predominantly highlight the (construct) validity issues of existing benchmarks across multiple task types including reasoning, bias, safety, and factuality (Raji et al., 2021; Salaudeen et al., 2025). Thus ideally, newer, more representative benchmarks are developed to address the construct validity issues of older benchmarks. However, construct validity issues are inherent to the benchmark format (Eriksson et al., 2025) and thus are unavoidable. In the best case, better benchmarks can be developed based on best practices cultivated from studying preceding implementations. However, this pathway grounded in continual benchmark development proves to be a laborious effort, often with less utility than desired as models quickly outscale and render benchmarks months old already saturated to at or above human baselines (Ott et al., 2022). This subsequently leads to a significant amount of epistemic waste, where these evaluations are left behind since they are no longer valuable towards tracking progress towards a high-level task construct.

Our solution is benchmancy, a new framework to recycle existing benchmarks by extending their epistemic validity ex post facto. In the scientific sense, benchmancy transforms benchmarking studies that only serve as a measuring stick for model performance into artifacts that can shed insight into the phenomenology of AI systems. Instead of building a whole new evaluation to capture a better representation of a construct, benchmancy adjusts the scope of the original benchmark. The evaluation suite of the original benchmark remains unchanged; in each run, the model is asked the same queries. What changes is the evaluation harness, the infrastructure that represents all of the instructions and tools necessary to run a benchmark end-to-end. If an evaluation harness represents the technical framing of the original aim of a benchmark, benchmancy provides a new framing to the tasks of the original benchmark in order to evaluate

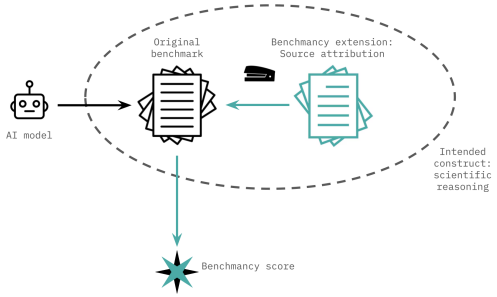

Figure 1: Illustration of the benchmancy method

an extension or broadening of the original construct. Specifically, we modify the system prompt to additionally request data on our construct extension; in this case, source attribution. The analysis stage is consequently also adjusted to now output a new, union score that represents both the original and benchmancy evaluation harnesses in order to provide a more accurate approximation of the intended construct, which in our case is the capacity for trustworthy scientific reasoning. Figure 1 outlines the benchmancy process for our example of scientific reasoning. Benchmancy is not designed to replace the need for newer, more valid benchmarks, but to provide a first pass at understanding the limitations of the original evaluation and expanding it in a fashion that is more representative of the real-world.

2.1 Scientific Reasoning

For AI model developers, science proves to be a rich wellspring of diverse, complex computational challenges to quantitatively prove their model's dominance over competitors. The value of science stems from its inherent epistemic integrity, a trait derived from its mandate for holistic measurability (Tal, 2020). As an epistemic system, science draws the line that everything that is science should be measurable and what is not measurable is not science. In the former case, every component of a scientific result has to be measurable. Importantly, measurement of any phenomenon does not imply complete comprehension as measurement itself does not imply complete precision. To be considered scientific, simply accomplishing a task is insufficient, the task must be completed in a provably accurate way. It is this level of challenging rigor, combined with the dense system of prior knowledge, that attracts developers to build "AI that doesn't just facilitate science but begins to do

science" (Huckins). In the adjacent field of mathematics, AI systems have already demonstrated their potential to work on research-level problems, generating a proof for a geometry problem originally posed by Paul Erdős in 1946 (OpenAI, 2026). For many scientists, the possibility of automating scientific research is believed to be an enticing pathway to magnify their research output, ultimately aiming to accelerate scientific understanding as a whole. In a parallel strain to full automation, works such as AlphaGenome (Avsec et al., 2026), AlphaEarth (team, 2025), and MatterGen (Zeni et al., 2025) demonstrate the power of AI tools trained to solve specific scientific tasks.

Efforts such as Sakana AI's AI Scientist (Yamada et al., 2025) and DeepMind's Co-Scientist (DeepMind, 2026) showcase potential futures for automated AI systems integrated into the scientific community. In the former's case, success was measured by determining if the output manuscript of their AI Scientist could be accepted at top-tier settings such as NeurIPS. While such a test is able to demonstrate the potential power of their workflow, it does not serve as a rigorous evaluation of the agent's ability to perform accurate scientific research. Without being able to measure underlying phenomena rather than external goalposts, it is impossible to ascertain to what degree an AI system is capable. Operating with this rationale in mind, a popular direction for benchmark developers recently is to ground their evaluations in more fundamental constructs such as scientific reasoning to assess extrapolative potential. By its nature, scientific reasoning is ubiquitous across the scientific research workflow, representing the systematic logic necessary to behave scientifically. Contemporary scientific reasoning benchmarks can be categorized into two classes: the first targeting domain-specific reasoning and the second, general-purpose (scientific) reasoning. Benchmarks such as ChemBench (chemistry) (Mirza et al., 2024), CritPt (physics) (Zhu et al., 2025), and MedQA (medicine) (Jin et al., 2020) aim to probe the capacity of AI systems to solve specific reasoning tasks constrained to a single field or sub-field of science. On the other end, benchmarks such as Humanity's Last Exam (Phan et al., 2025), FrontierScience (**?**), GPQA (Rein et al., 2023), and MMLU (Hendrycks et al., 2021) attempt to evaluate AI performance on a variety of reasoning tasks sampled from several fields. For our study, we will be focusing on the latter class of scientific reasoning benchmarks.

The test datasets of the aforementioned scientific reasoning benchmarks are composed of isolated, expert-written questions, sometimes with multiple choice responses as with MMLU (Hendrycks et al., 2021) and GPQA (Rein et al., 2023). The questions challenge models to factually solve complex, single-turn problems. However, this task construction is incomplete in its ability to represent scientific reasoning "in the wild." In their launch of FrontierScience, OpenAI describes that scientists already use AI models "to accelerate research workflows while relying on human judgment for problem framing and validation, and increasingly to explore ideas and connections that would otherwise take much longer to uncover—including, in some cases, contributing new insights that experts then evaluate and test" (Wang et al.). In practice, scientific research is significantly less orderly with tasks occurring simultaneously with interconnected dependencies. Tasks can no longer be viewed as independent, but are embedded in an evolving structure that results in scientific discovery (Latour, 1999). Assessments of factuality and logical operation are therefore only a part of what constitutes scientific reasoning. Valid source attribution or engagement with the existing compendium of scientific knowledge is equally as critical, yet is absent from scientific reasoning benchmarks. Scientists have already been actively discussing the negative ramifications incurred by hallucinated references (Sharifi, 2025; Jones, 2025; Agrawal et al., 2024; Walters and Wilder, 2023). Even focusing on citations alone already reveals how AI systems can degrade the epistemic integrity of science. However, citations are but one channel by which scientists engage with the knowledge ecosystem of science. Subtler, but no less dangerous emergent harms can arise beyond simple citation hallucination. Thus, we will use our new framework of benchmancy to study a more construct-valid broadening of scientific reasoning that includes source attribution.

3 Experimental Design

As illustrated by Figure 1, in order to study the expanded construct of scientific reasoning inclusive of source attribution, we select four notable academic or scientific reasoning benchmarks: MMLU (Hendrycks et al., 2021), GPQA (extended version) (Rein et al., 2023), Humanity's Last Exam (Phan et al., 2025), and FrontierScience (Wang et al.). These benchmarks were chosen due to their broad

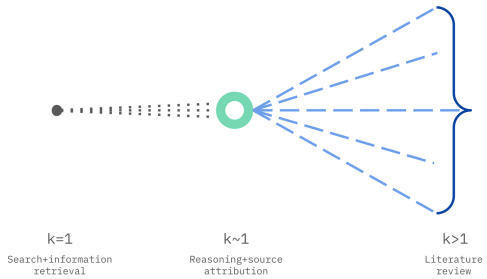

Figure 2: Scope of attributional tasks with respect to the number of references expected (k)

coverage of multiple domains of science and stated question complexity at or above the collegiate level. For each benchmark, we only selected queries that fell into the following categories (if present): biology, chemistry, computer science, engineering, and physics. After the subset selection, there are 1,027 queries for Humanity's Last Exam, 546 queries for GPQA, 619 queries for MMLU, and 160 queries for FrontierScience.

To perform benchmancy on the aforementioned benchmarks, we adjusted the system prompt to additionally request at least one academic article that would represent the most relevant reference to its response generation. Each reference should contain a title, author list, year of publication, and open-access url link. The exact system prompt used for our studies can be found in Appendix A. While there is existing literature evaluating LLMs on their ability to perform search and information retrieval tasks or even assist in literature reviews (Press et al., 2024; Chen et al., 2025; Bragg et al., 2025; Xiong et al., 2026), by choosing to let the model output at least one reference our study seeks to investigate an understudied area in terms of source attribution. The agency provided to the model mimics what is done by human scientists when solving scientific reasoning queries, where the requisite knowledge is sourced from a limited number of references (which we pose is on the order of one). Section **??** further expands how our choice creates a unique niche for our work among contemporary literature. Figure 2 identifies the critical region we are targeting, situated between information retrieval tasks that try to find one specific reference and literature reviews that request multiple to explore a concept. Scientists frequently use AI systems as reasoning assistants and thus will ask complex, single-turn queries and expect scientific rigor in the models' responses. Through benchmancy, we can evaluate the performance of models on this form of source attribution that is common in the user space, but lacks any forms of evaluation. Additionally, Berman et al. (2026) has shown that LLMs can exhibit suppression of attribution as a unique failure mode even when the information is available. By not strictly defining how many references we desire, we are then able to assess a model's propensity of suppression when attached to a scientific reasoning query.

For each benchmark, we follow the original evaluation harness in calculating the original performance score for scientific reasoning. In the case of MMLU and GPQA, since these are multiple-choice benchmarks, we directly check the outputs. For Humanity's Last Exam and FrontierScience, we use deepseek-v4-flash (DeepSeek-AI, 2026) as a judge LLM to grade the accuracy of the tested model's response. The benchmancy method is tested using four LLMs: Gemma4 (effective 4B) (DeepMind), Mistral-7B (Jiang et al., 2023), GPT-5-mini (Singh et al., 2025), and Claude Haiku-4.5 (Anthropic). Each pass of the benchmancy harness is performed five times to account for statistical fluctuations. Finally, we calculate the benchmancy score, which is representative of the model's performance on the expanded task of scientific reasoning with source attribution:

$$x_B = x_\alpha \times x_c \times s_m \tag{1}$$

where:

x_B = The overall score output from the benchmancy harness,

x_α = The accuracy score of the original benchmark,

x_c = The accuracy of all generated citations,

s_m = The output of Semantic Scholar's relevance function to calculate the match score between the provided title and the found reference

4 Results

Figure 3 depicts the accuracy scores for each "benchmanced" benchmark for all tested models. As described in Equation 1, the benchmancy accuracy is calculated as the product of the accuracy on the original benchmark (x_α, original accuracy), the binary score of whether the generated title matches an entry in Semantic Scholar (Kinney et al., 2023) (x_c, citation accuracy), and the match score derived from Semantic Scholar's relevance function that

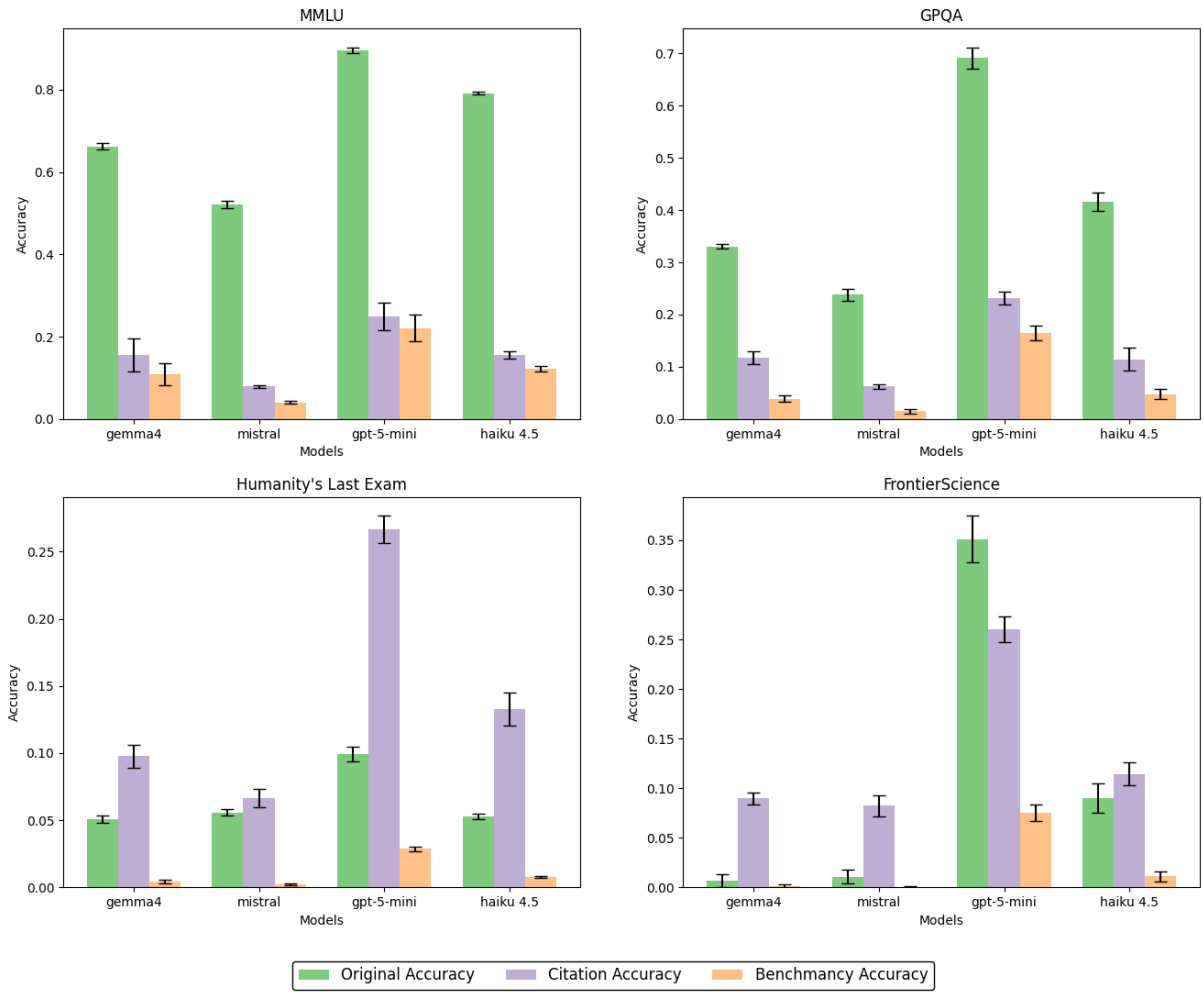

Figure 3: Overall scores for the benchmancy harness

encodes how well the generated title connects to Semantic Scholar's matched reference (s_m).

Over all evaluation runs, GPT-5-mini was the highest performing model. On MMLU and GPQA, all models performed significantly better on the original benchmark task than the source attribution extension. The opposite is true for Humanity's Last Exam and FrontierScience, with the exception of GPT-5-mini that had a higher original accuracy on FrontierScience in contrast to its citation accuracy. While GPT-5-mini outperformed the expert human baseline of 64.8% on GPQA (Rein et al., 2023), the addition of the benchmancy harness lowered performance significantly with a benchmancy accuracy of 16.5%. In this manner, GPQA, which was a saturated benchmark for GPT-5-mini, now has reclaimed utility due to our ex post facto enhancement.

While the accuracy scores provide a coarse look at how the addition of source attribution affects model performance on the selected benchmarks,

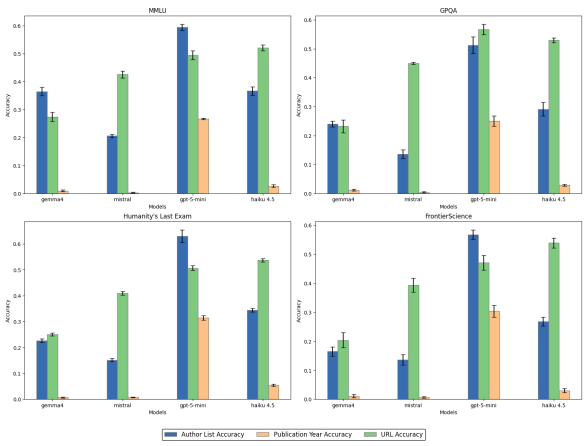

Figure 4: Accuracy scores for the metadata fields of the generated citations

we can further break down the causes for this reduction by looking at how well the models were able to provide the correct metadata fields for a generated reference. It should be noted that the citation accuracy as shown in Figure 3 only counts

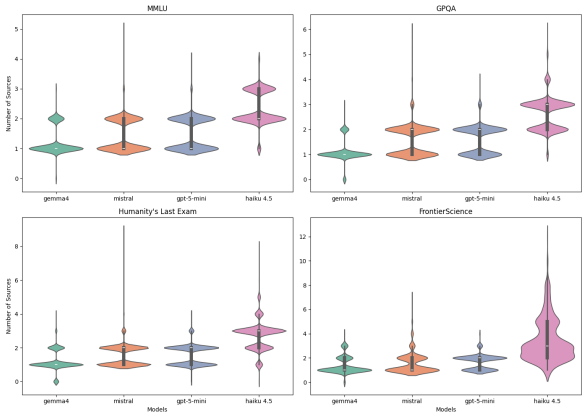

Figure 5: Distributions of the number of generated sources

how many instances a generated title yields a valid Semantic Scholar entry. Figure 4 shows the accuracy of the other metadata fields requested by our benchmancy harness: author list, year of publication, and (open-access url). The denominator in Figure 4 is equal to the numerator of citation accuracy from Figure 3 or, in other words, when the title produces a valid reference by Semantic Scholar. Author list accuracy corresponds to the model producing at least one correct author for a given source. URL accuracy refers to the model outputting a valid URL. Appendix B expands on this by comparing two different URL accuracies: 1) exact matching of URLs between generated reference and Semantic Scholar's entry and 2) the URL being real, which is what is being called URL accuracy in Figure 4. We find that all models for all benchmancy instances fail to produce the exact URL logged by Semantic Scholar. However, this is very possibly the result of numerous open-access URLs mapping to the same source.

Models had consistent performance trends across benchmarks with the ranking of accuracies staying relatively consistent. For example, both Mistral and Claude Haiku 4.5 performed best on URL accuracy, followed by author list accuracy, and URL accuracy last. Overall, the worst performing metadata field was publication year accuracy. Despite the lenient definitions for author list and URL accuracy, all models generated a valid author within an author list or a valid URL generally less than 50% of the time. Even if the model produced a valid title, many corresponding metadata fields were hallucinated nonsense with authors such as "First Name Last Name," publication years such as "2XXX," and URLs such as "example.com."

We can also look at the distribution of the number of references generated across all benchmancy instances to quantify how much support a model deems necessary for a scientific reasoning query. Figure 5 shows the distribution of the number of references generated across all benchmancy instances. We recall that as per our system prompt (Appendix A), the model should always output at least one source. Figure 5 proves that this is generally true with all models showing average source counts at or above one. Claude Haiku 4.5 is the most rigorous in this capacity with an average source count of 2.8. The distribution presented in Figure 5 can point to a critical epistemic failure mode known as suppression in which the model deliberately omits attribution (Berman et al., 2026). Gemma4 consistently demonstrates non-zero counts of suppression across all benchmancy instances. GPT-5-mini and Claude Haiku 4.5 only show suppression for Humanity's Last Exam.

5 Related Works

5.1 Repurposing Benchmarks

The power of benchmancy lies in its simplicity. The method is able to extend the construct validity of existing benchmarks without significantly changing their evaluation harness. However, the idea of repurposing benchmarks as raw datasets is not new and has been implemented in various forms since 2024. Park et al. (2025) used popular vision-language benchmarks such as MMStar and RealWorldQA to detect contaminated VLMs and show that they would fail to generalize under controlled semantic perturbations. Li et al. (2024) built a dataset extension of GSM8K known as GSM-Plus that tested the robustness of LLMs on mathematical reasoning tasks by perturbing the queries of the former benchmark. Golchin and Surdeanu (2023) devises the Data Contamination Quiz, which is able to detect verbatim data contamination in LLMs for popular benchmarks such as RTE, AGNews, and XSum by creating quiz questions with perturbed versions of a target query. Unlike these works, benchmancy is able to probe a related task to the original benchmark's evaluation target with minimal perturbations to its construction.

5.2 Source Attribution

Accurate knowledge traceability is a computational aim that stems from pre-AI research on search, information retrieval, and knowledge discovery in

databases (KDD) and now has translated into the space of LLM evaluations. The past year has witnessed several research endeavors into uncovering how LLMs struggle with accurate citation generation (GPTZero, 2025). Efforts such as AttriBench extend this investigative push to the broader task of source attribution, showing that LLMs cannot accurately attribute (or even obscure attribution of) content to its original authors (Berman et al., 2026). However, AttriBench only focuses on general scholastic questions and does not engage with scientific literature. In addressing scientific source attribution, the literature is still limited. The CiteME benchmark recovers masked in-text citations (Press et al., 2024) and Pub-Guard-LLM predicts the likelihood of a paper to be retracted based on methodological flaws, metadata inconsistencies, etc. for biomedical articles (Chen et al., 2025). ELAIPBench shows that LLMs perform far below human performance on comprehending AI research papers (Dai et al., 2025). Through its scientific benchmarking suite AstaBench, Ai2 released several literature understanding benchmarks with a variety of aims ranging from testing agents on their ability to locate sets of papers given a natural language description (PaperFindingBench) to generating deep-research-style reports for CS literature review questions (ScholarQA-CS2) (Bragg et al., 2025). As illustrated in Figure 2, our study targets a different task niche of providing relevant references for complex, single-turn reasoning queries.

6 Conclusion

Etymologically, benchmancy is derived from the Greek *manteia*, roughly translating to divination. Unlike divination from soil (geomancy) or the spirits of the dead (necromancy), we seek to "divine" or gain insight from the existing landscape of AI benchmarks. Our benchmancy studies into expanding the construct of scientific reasoning to include source attribution shows that frontier LLMs are unable to perform adequately on this more realistic enhancement of the original tasks laid out in the investigated benchmarks. Additionally, we show that this reduction in performance parallel to task broadening reclaims previously saturated benchmarks. This provides new opportunities to reclaim older benchmarks as sources of queries that can be used to probe new phenomena about their evaluation target or construct.

It should be noted that benchmancy is not designed to be the solution to construct validity issues with benchmarks, but a first-pass approach to provide valuable information that can inform the creation of a more holistic evaluation in the future. It is our hope that our work on benchmancy inspires the creation of an ecosystem of evaluations with varying stakes with benchmancy being an efficient, low-stakes probe into the epistemic validity of existing benchmarks.

7 Ethics Statement

AI assistance (Gemini 3.1 Pro, Antigravity) was used to assist with the coding for the experiment described in this publication.

Limitations

Benchmancy as a method of enhancing the epistemic validity of existing benchmarks is inherently limited. Most obviously, since the queries from the original benchmark are used exactly, there is an inheritance of any construct validity issues embedded by the query design. Additionally, benchmancy is only applicable when trying to augment a construct such as scientific reasoning with a related task that co-occurs in the real-world such as source attribution.

Our experiment of expanding scientific reasoning benchmarks to include source attribution is currently only limited to assessing LLMs out of the box. Thus, we are not testing any additional tools such as RAG, agentic harnesses, or web search that might significantly boost the source attribution accuracy.

Both MMLU and Humanity's Last Exam originally contain more queries than what was used in our experiment. To ensure equitable comparisons and constrain the evaluation construct as scientific reasoning, we only chose questions pertaining to biology, chemistry, computer science, engineering, and physics. However, this constitutes a significant alteration to the original benchmark.

For our experiment, we prompted the models to only output academic articles as references. However, the knowledge necessary to solve scientific reasoning queries can also easily be found in books, blog posts, or other web-hosted sources. To verify the generated articles, we relied on the Semantic Scholar API. However, the Semantic Scholar database contains many entries with incomplete metadata, which negatively skews performance on metadata accuracy in a manner that is not related to

the benchmancy harness. Additionally, the manner by which the URL metadata field was checked was also limited, either by just checking if the URL is real as in Figure 4 or through direct matching with what is contained in Semantic Scholar's database (Appendix B).

References

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Tauman Kalai. 2024. Do Language Models Know When They're Hallucinating References? *arXiv preprint*. ArXiv:2305.18248 [cs].

Anthropic. Claude Haiku 4.5.

Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R. Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, Raina Thomas, Vincent Dutordoir, Matteo Perino, Soham De, Alexander Karollus, Adam Gayoso, Toby Sargeant, Anne Mottram, Lai Hong Wong, and 8 others. 2026. Advancing regulatory variant effect prediction with AlphaGenome. *Nature*, 649(8099):1206–1218.

Eliza Berman, Bella Chang, Daniel B. Neill, and Emily Black. 2026. Attribution Bias in Large Language Models. *arXiv preprint*. ArXiv:2604.05224 [cs.AI].

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi, Sergey Feldman, Dany Haddad, Jena D. Hwang, Peter Jansen, Varsha Kishore, Bodhisattwa Prasad Majumder, Aakanksha Naik, Sigal Rahamimov, Kyle Richardson, Amanpreet Singh, Harshit Surana, Aryeh Tiktinsky, Rosni Vasu, Guy Wiener, and 20 others. 2025. AstaBench: Rigorous Benchmarking of AI Agents with a Scientific Research Suite.

Lihu Chen, Shuojie Fu, Gabriel Freedman, Cemre Zor, Guy Martin, James Kinross, Uddhav Vaghela, Ovidiu Serban, and Francesca Toni. 2025. PubGuard-LLM: Detecting Retracted Biomedical Articles with Reliable Explanations. *arXiv preprint*. ArXiv:2502.15429 [cs].

Xinbang Dai, Huikang Hu, Yongrui Chen, Jiaqi Li, Rihui Jin, Yuyang Zhang, Xiaoguang Li, Lifeng Shang, and Guilin Qi. 2025. ELAIPBench: A Benchmark for Expert-Level Artificial Intelligence Paper Understanding.

DeepMind. Gemma 4.

DeepMind. 2026. Co-Scientist: A multi-agent AI partner to accelerate research. Section: Science.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Maria Eriksson, Erasmo Purificato, Arman Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. 2025. Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation.

Shahriar Golchin and Mihai Surdeanu. 2023. Data Contamination Quiz: A Tool to Detect and Estimate Contamination in Large Language Models.

GPTZero. 2025. GPTZero uncovers 50+ Hallucinations in ICLR 2026.

GPTZero. 2026. GPTZero finds 100 new hallucinations in NeurIPS 2025 accepted papers.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *arXiv preprint*. ArXiv:2009.03300 [cs].

Grace Huckins. Google I/O showed how the path for AI-driven science is shifting.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *arXiv preprint*. ArXiv:2009.13081 [cs].

Nicola Jones. 2025. AI hallucinations can't be stopped — but these techniques can limit their damage. *Nature*, 637(8047):778–780. Bandiera_abtest: a Cg_type: News Feature Subject_term: Machine learning, Computer science, Technology.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David Graham, Fangzhou Hu, and 29 others. 2023. The Semantic Scholar Open Data Platform. Version Number: 2.

B. Latour. 1999. *Pandora's Hope: Essays on the Reality of Science Studies*. Pandora's Hope: Essays on the Reality of Science Studies. Harvard University Press.

Mingtang Li, Giacomo Livan, and Simone Righi. 2025. Quantifying the dynamics of peak disruption in scientific careers. *Scientific Reports*, 15(1):10812. Publisher: Nature Publishing Group.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. 2024. GSM-Plus: A Comprehensive Benchmark for Evaluating the Robustness of LLMs as Mathematical Problem Solvers. *arXiv preprint*. ArXiv:2402.19255 [cs.CL].

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, Amir Mohammad Elahi, Mehrdad Asgari, Juliane Eberhardt, Hani M. Elbeheiry, María Victoria Gil, Maximilian Greiner, Caroline T. Holick, Christina Glaubitz, Tim Hoffmann, and 16 others. 2024. Are large language models superhuman chemists? *arXiv preprint*. ArXiv:2404.01475 [cs].

Miryam Naddaf. 2026. How much of the scientific literature is generated by AI? *Nature*. Bandiera_abtest: a Cg_type: News Subject_term: Computer science, Peer review, Technology.

OpenAI. 2026. An OpenAI model has disproved a central conjecture in discrete geometry.

Simon Ott, Adriano Barbosa-Silva, Kathrin Blagec, Jan Brauner, and Matthias Samwald. 2022. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1):6793. ArXiv:2203.04592 [cs].

Jaden Park, Mu Cai, Feng Yao, Jingbo Shang, Soochahn Lee, and Yong Jae Lee. 2025. Contamination Detection for VLMs using Multi-Modal Semantic Perturbation.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, and 1093 others. 2025. Humanity's Last Exam. *arXiv preprint*. ArXiv:2501.14249 [cs].

Ori Press, Andreas Hochlehnert, Ameya Prabhu, Vishaal Udandarao, Ofir Press, and Matthias Bethge. 2024. CiteME: Can Language Models Accurately Cite Scientific Claims? *arXiv preprint*. ArXiv:2407.12861 [cs].

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the Everything in the Whole Wide World Benchmark. *arXiv preprint*. ArXiv:2111.15366 [cs].

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint*. ArXiv:2311.12022 [cs].

Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. 2025. Measurement to Meaning: A Validity-Centered Framework for AI Evaluation. *arXiv preprint*. ArXiv:2505.10573 [cs] version: 4.

Ayyoob Sharifi. 2025. Tackle fake citations generated by AI. *Nature*, 644(8075):42–42. Bandiera_abtest: a Cg_type: Correspondence Subject_term: Research management, Machine learning.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, A. J. Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2025. OpenAI GPT-5 System Card.

Nathan Suri and Savannah Jennifer Thais. 2025. Holonic science: A new framework for benchmarking AI scientists. In *NeurIPS 2025 AI for Science Workshop*.

Eran Tal. 2020. Measurement in Science. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, fall 2020 edition. Metaphysics Research Lab, Stanford University.

The AlphaEarth Foundations team. 2025. AlphaEarth Foundations helps map our planet in unprecedented detail. Section: Science.

Thomas G. Dietterich [@tdietterich]. 2026. Attention @arxiv authors: Our Code of Conduct states that by signing your name as an author of a paper, each author takes full responsibility for all its contents, irrespective of how the contents were generated. 1/.

William H. Walters and Esther Isabelle Wilder. 2023. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13(1):14045.

Miles Wang, Robi Lin, Kat Hu, Joy Jiao, Neil Chowdhury, Ethan Chang, and Tejal Patwardhan. FrontierScience: Evaluating AI's Ability to Perform Expert-Level Scientific Tasks.

Lei Xiong, Kun Luo, Ziyi Xia, Wenbo Zhang, Jin-Ge Yao, Zheng Liu, Jingying Shao, Jianlyu Chen, Hongjin Qian, Xi Yang, Qian Yu, Hao Li, Chen Yue, Xiaan Du, Yuyang Wang, Yesheng Liu, Haiyu Xu, and Zhicheng Dou. 2026. AutoResearchBench: Benchmarking AI Agents on Complex Scientific Literature Discovery.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search. *arXiv preprint*. ArXiv:2504.08066 [cs].

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, and 7 others. 2025. A generative model for inorganic materials design. *Nature*, 639(8055):624–632.

Zhenyue Zhao, Yihe Wang, Toby Stuart, Mathijs De Vaan, Paul Ginsparg, and Yian Yin. 2026. LLM hallucinations in the wild: Large-scale evidence from non-existent citations. *arXiv preprint*. ArXiv:2605.07723 [cs.DL].

Minhui Zhu, Minyang Tian, Xiaocheng Yang, Tianci Zhou, Lifan Yuan, Penghao Zhu, Eli Chertkov, Shengyan Liu, Yufeng Du, Ziming Ji, Indranil Das, Junyi Cao, Yufeng Du, Jiabin Yu, Peixue Wu, Jinchen He, Yifan Su, Yikun Jiang, Yujie Zhang, and 45 others. 2025. Probing the Critical Point (CritPt) of AI Reasoning: a Frontier Physics Research Benchmark. *arXiv preprint*. ArXiv:2509.26574 [cs].

A Benchmancy System Prompt

""" Your response must be a valid JSON object with the following exact structure. Provide AT LEAST ONE academic article in the references array for the most relevant academic reference(s) that your answer is based on. You may include multiple references if needed. Ensure all author names are formatted as "First Name Last Name".

"answer": "your chosen answer", "references": ["title": "title of article", "authors": ["First Last", "First Last", "... (include all authors formatted as First Name Last Name)"], "date": "date of publication for article", "url": "open-access web link to article" , "...": "include additional reference objects here if needed"] """

B URL Accuracy

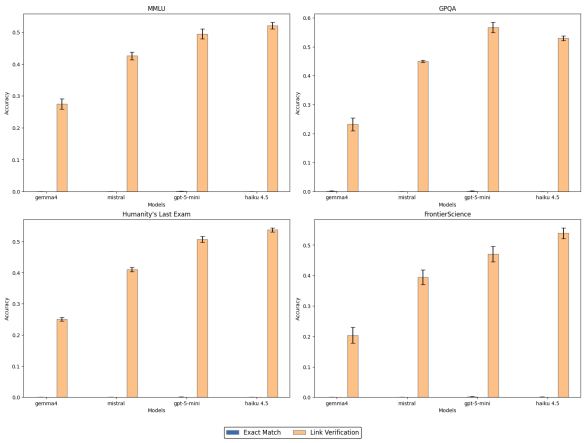

Figure 6: Comparison between URL matching and verification accuracies