MULTIFACETED EVALUATION OF LARGE LANGUAGE MODELS' POTENTIAL FOR SCIENTIFIC DISCOVERY

Authors: Olivia Barsemian, Sophie Chen, XiaoQing Cui, Sahasra Datla, Daniel Egusquiza Bronstein, Sreethan Gangavarapu, Justin Lambert, John Loatman, Nathaniel Nadler, Shivan Shrinivas, Michael Xu, Adhithya Yuvaraj

Advisor: Dr. Minjoon Kouh

Assistants: John McLaughlin, Giorgi Tsnobiladze

ABSTRACT

Large language models (LLMs) have become increasingly popular in scientific research, yet their ability to autonomously perform each individual task involved in scientific discovery remains uncertain. We developed four proof-of-concept benchmarks that evaluated LLMs' abilities in recognizing patterns in integer sequences, understanding whether given information is sufficient to make a conclusion, identifying human errors in lab reports, and deducing implications of alternative physical laws. We identify the disconnect between model accuracy and confidence and the consequences of model failure as key considerations when integrating LLMs into research processes. Ultimately, we conclude that the central concern of LLM benchmarking in a scientific discovery context is not determining whether models can serve as a scientist, but rather discerning which components of scientific reasoning can be performed reliably, the conditions under which they can be performed, and the areas in which failures occur.

INTRODUCTION

While researchers disagree over whether LLMs are a fad or a century-defining invention, their rapid integration into all stages of the scientific process is a fact. To assess the quality of this surge of LLM output, researchers have developed benchmarks, or structured sets of evaluation tasks assessing frontier models across a variety of scientific tasks (Zhao 2026). For example, Humanity's Last Exam (Center for AI Safety et al. 2026) tests in-depth reasoning using expert-level questions with established answers. Another set of examples is the FrontierScience (Wang M et al. 2026) and Scientific Discovery Evaluation benchmarks, (Zhangde et al. 2026) which attempt to assess models on open-ended scientific questions. Systems such as the AI Scientist (Lu, C et al. 2026) go further by using LLMs to automate ideation, experimentation, analysis, and paper writing. However, the mechanisms by which LLMs are able to scientifically reason are still poorly understood, which raises doubts about whether current LLMs can effectively contribute to scientific research.

One factor which makes it difficult to guarantee the quality of LLM-generated science is that LLMs do not reason through language in the same manner as humans. At their core, LLMs were designed as sophisticated "word-predictors," not scientists. They train by consuming immense amounts of human-written text, then breaking the text up into "tokens." One example could be breaking up "incomprehensible" into the tokens "in," "comprehens," and "ible." Through a mechanism known as the "transformer" (Vaswani et al. 2017), LLMs figure out which tokens are most highly correlated and therefore which tokens are the most "important." Using its knowledge of importance, the model is able to predict the next token from a given prompt. Although this process often produces coherent outputs, it can also confidently generate plausible yet incorrect claims, commonly described as "hallucinations" (Kumaran et al. 2026). This

tendency raises concerns about the reliability and accuracy of LLM-generated scientific reasoning and conclusions (Zhao et al. 2026; Naveed et al. 2025). Therefore, evaluating scientific capacity requires qualitatively examining how models recognize patterns, evaluate evidence, identify methodological errors, and reason under unfamiliar assumptions.

To this end, we develop proof-of-concept benchmarks that evaluate distinct capabilities involved in scientific discovery, as shown in Figure 1. Specifically, we implemented four benchmarks, each intended to address a distinct element of the scientific method (Hepburn and Andersen 2015). The first benchmark tests pattern recognition and numerical reasoning by prompting models to extend mathematical sequences and predict a general rule. The second examines whether models can gauge if they have sufficient evidence to support an underlying trend and draw appropriately cautious conclusions from numerical data. The third evaluates models' ability to identify errors in laboratory reports, and similarly gauge if those errors can be proven. The fourth benchmark tests the extent to which LLMs can push the boundaries of science by asking them to identify alternative hypothetical physical laws, then generate hypotheses and design experiments in order to explore the implications of those laws.

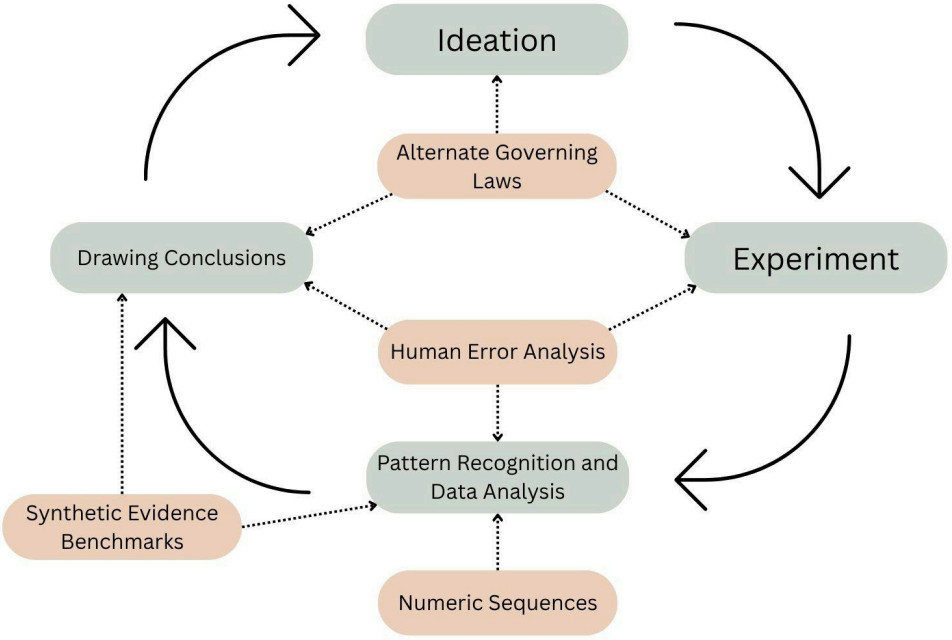


Figure 1. Mapping of the four benchmarks onto key stages of the scientific method.

GENERAL METHODOLOGY

Several components of our methodology were shared between benchmarks. We used ChatGPT 5.6 Sol, Claude Opus 4.8, and Gemini 3.6 Thinking in our trials to survey a variety of frontier models. In general, when designing the benchmarks, we tried to ensure that the tasks we assigned were neither too easy nor too unreasonable for the LLMs to complete, and included a range of difficulties within the benchmarks. This allowed us to observe changes in performance across tasks and difficulties. We avoided tasks that revolved around solely factual recall or information that could be easily obtained with a search engine because we wanted to isolate

scientific reasoning and discovery. This was achieved by creating original prompts that instructed LLMs to not use internet searches. However, this instruction does not definitively guarantee that no searches were performed. Finally, we prevented LLMs from using previous context by disabling memory, using built-in incognito functions, “temporary chat” features, and isolated accounts and different prompt branches.

Our studies also possessed several limitations which constrained the scope and applicability of our results. These experiments were generally performed with small sets of responses: generally, around 10-20 prompts were used in each lab. This means that our findings should be interpreted as proofs-of-concept for our specific style of benchmark, rather than a conclusive statement regarding LLMs’ scientific capabilities. Since each benchmark tests a single skill within the scientific method, any results reflect the model's ability to reason in that particular skill, not a general capacity across the scientific method. Finally, we used consumer interfaces for our testing, meaning that we could not measure hidden parameters like system prompts, with the benefit of more accurately replicating the way most people actually use LLMs in their work.

PART A - NUMERIC PATTERN RECOGNITION

Methods

In this benchmark, we evaluated LLMs’ ability to recognize the underlying pattern from a given sequence and generate the next two terms. The following is a description of each type of sequence we used to prompt the LLMs. Representative examples of each are provided in Appendix A; some sequences are sourced from the Online Encyclopedia of Integer Sequences (OEIS Foundation Inc. 2026).

1. Closed-form sequences generate each term directly from its position using a formula; they test a model’s ability to infer explicit mathematical relationships from limited data.

2. Recursive sequences generate each term using one or more preceding terms; they test LLMs’ ability to identify sequential dependencies.

3. Number-theoretic sequences consist of integers that possess common properties, such as primality and divisibility; they test whether LLMs can recognize numerical properties beyond simple differences and ratios.

4. Combinatorial sequences, including counting arrangements and paths, assess the LLM’s ability to determine and perform complicated counting-based calculations.

5. Constant category sequences form from digits of mathematical constants such as π and e; they test knowledge retrieval and numerical recognition skills of LLMs.

6. Specialized mathematical sequences, such as Babylonian Spiral, are generated using rules from specific areas of mathematics, including geometry and game theory; they test whether LLMs could apply specialized knowledge.

We crafted 24 benchmark questions; each of which was based on distinct sequences, with four sequences in each of the predefined categories. Each sequence was tested twice to soften the effect of prompt-to-prompt variations and increase the variety of our results. A standardized

prompt was used in each benchmark, which can be found in Appendix B. In short, each prompt included the first six terms of the sequence, asked the LLM to return solely the next two elements and the general rule, and prompted the models to not search the internet.

Results

During the trials, output from the models were classified as either identifying the intended sequence, identifying a valid alternative, or failing to identify a sequence consistent with the supplied data. These were separated into the groups displayed in Table I, where the valid reasoning rate was defined as the sum of the intended and alternative, as it represents the percentage of mathematically sound responses.

Table I. Performance of GPT-5.6 Sol High and Claude Opus 4.8 high against sequences. Note that models proposed alternate sequences fairly often, and had high accuracy rates overall.

Metric	ChatGPT	Claude
Intended	44%	63%
Alternative	25%	23%
Incorrect	31%	14%
Valid Reasoning Rate	69%	86%

Performance differed among the six mathematical categories as shown in Figure 2. Closed-form/explicit sequences received the highest success rate among the four tests, suggesting that both models were generally effective at inferring direct functional relationships from a limited amount of data. In contrast, constant-based sequences produced the lowest overall accuracy, as they often lacked obvious numerical or formulable patterns, making it more difficult for the LLMs to recognize specific windows of certain mathematical constants. Please note that since each sequence category only had four sequences, these category-level comparisons should be treated as a descriptive and exploratory study rather than definitive evidence of performance difference of and between models.

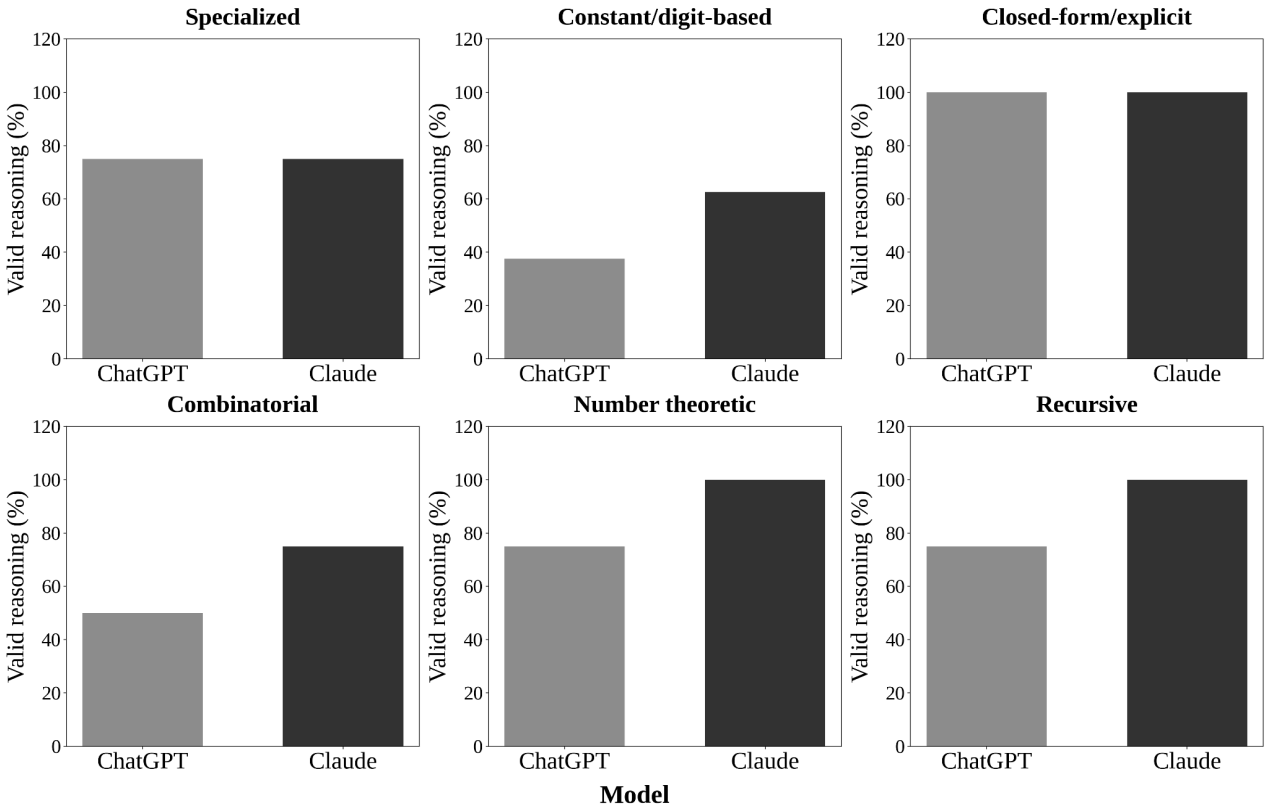


Figure 2. Valid reasoning rate by sequence category across both trials. Both models performed best on explicit sequences and worst on constant sequences.

Discussion

This benchmark examined an important aspect of scientific reasoning: drawing conclusions from limited observations. The fact that both models performed best on explicit sequences and worst on constant sequences suggests that LLMs are more effective at identifying clear mathematical relationships than recognizing more abstract patterns in data. This could limit their suitability for autonomous, high-level data analysis when the goal is to uncover the most insightful relationships within the data, which may not always be obvious. In a broader context, the frequency of models producing alternative sequences highlights an important limitation of using LLMs for scientific discovery: they can generate plausible explanations that fit the data, but not necessarily the explanation the researcher is looking for. Thus, LLMs can be useful for scientific tasks like hypothesis generation and data analysis, but they should still be used with caution and high human oversight. However, these conclusions remain limited by the small sample size, which prevented us from conducting statistical analysis and may not span the wide spectrum of potential pattern types and sequence difficulties.

PART B - EVIDENTIAL STATE CLASSIFICATION

Methods

We crafted a benchmark testing whether models could distinguish among three scenarios: sufficient evidence for selecting one candidate hypothesis, resolvable uncertainty among multiple

plausible hypotheses given further experimentation, and failure of the supplied hypotheses to adequately explain the observations. We used unspecified mathematical systems as our hypotheses, as we wanted to control the LLMs' ability to rely on known scientific relations (e.g., the formula describing net force, $F = ma$, is widely available online and LLMs are likely trained on such knowledge).

Each prompt presented five data points (x_i, y_i) that showcased a relationship between a controlled input x and an observed response y. The model was then presented with three candidate hypotheses, represented by single-variable functions f, g, and h. Then, given these data points and hypothesized relationships, the model was prompted to decide between three scenarios: whether the evidence clearly supported one function; whether the current evidence was insufficient to make a decision, but additional experimentation could distinguish the correct hypothesis; or whether none of the proposed functions adequately fit the data. Then, we asked it to report its confidence as a percentage from 0 to 100.

When creating prompts, we used a variety of underlying functions including linear, polynomial, trigonometric, rational, absolute-value, sigmoidal, and combinations thereof. Each test case contained five input values (x values) drawn from the interval [-5, 5]. In order to justify our placement of any example into one of the decidability categories, we defined these categories numerically based on quantitative error metrics. Further details as to our exact mean squared error (MSE) thresholds and experimental setup are given in Appendix C.

Responses were either considered correct or incorrect, with one point assigned to correct classifications and zero points for incorrect classifications.

Finally, confusion matrices were used to examine not only how often each model was wrong, but also the direction of its errors. These 3x3 tables compare the true evidence scenario with the scenario selected by the model, revealing which situations were most often confused with one another (Figure 3).

Results

We used two main statistical methods: a technique known as the top-level Brier score, which measures the difference between a model's stated confidence and its observed correctness. Lower Brier scores indicate better agreement between how certain a model is and how often it is actually correct. We also used 95% Wilson confidence intervals that estimate a plausible range of values for the true accuracy of the models tested. These intervals specialize in small sample sizes and data solely between 0 and 1.

In 30 trials, both Claude Opus 4.8 and ChatGPT Pro assigned 21 benchmark items the correct evidentiary state, corresponding to an accuracy rate of 70.0%. The 95% Wilson confidence interval for each estimate ranged between 52.1% and 83.3%, reflecting the uncertainty associated with the relatively small sample size (Table II).

Performance varied considerably across the three evidentiary categories. Both models correctly identified 10 out 10 cases of model misspecification; they were highly reliable in identifying cases in which none of the proposed hypotheses was an accurate representation of the data provided. On the other hand, ChatGPT correctly identified 3 out of 10 cases in which there was sufficient evidence, 8 out of 10 cases in which there was resolvable uncertainty, while

Claude Opus 4.8 managed one more correct response in sufficient evidence, but one less correct response in resolvable uncertainty (Table II).

Table II. Summary statistics of model performance across all tested cases. Both models display a very similar performance but very different confidence in their responses.

Model	Trials	Correct	Accuracy	Confidence Interval		Mean Confidence	Brier Score
				Low	High		
Opus 4.8	30	21	0.70	0.521	0.833	0.834	0.216
ChatGPT Pro	30	21	0.70	0.521	0.833	0.947	0.279

These category level results show that aggregate accuracy alone is an incomplete assessment of LLM capacity. Although the models achieved identical overall scores (Table II), both differed in their confidences and therefore Brier scores. Claude obtained a lower top-label Brier score than ChatGPT, 0.216 compared with 0.279, indicating closer descriptive agreement between its confidence and observed correctness. ChatGPT remained highly confident even though nine of its 30 classifications were incorrect. It is notable that both models were much more successful at rejecting inadequate candidate mechanisms than concluding that the evidence sufficiently favored one hypothesis, suggesting excessive caution.

The most common way in which models failed was by prematurely rejecting all hypotheses in favor of concluding a model misspecification, as seen in Figure 3's rightmost columns on both models. However, they exhibited a 100% accuracy rate on classifying model misspecifications.

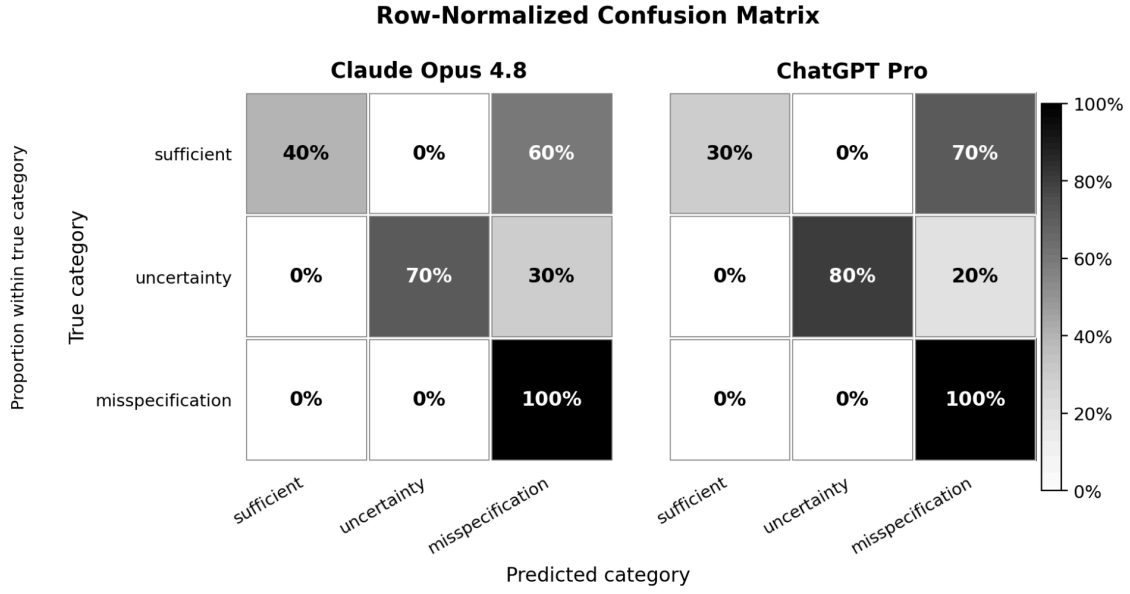


Figure 3. Confusion matrix for Claude Opus 4.8 (left) and ChatGPT Pro (right). Demonstrates LLM's predicted category for the true category and shows common mistakes and successes.

Discussion

This study suggests that frontier LLMs can distinguish whether they have enough information to make conclusions, but are predisposed to reject rather than confirm hypotheses. Both models reliably recognized cases in which none of the supplied hypotheses adequately represented the provided data, while also frequently extending that logic to cases in which one or more of the proposed mechanisms were sufficiently supported (see Figure 4 for an example). They demonstrate that there are also valid concerns that LLMs may exhibit a limited sense of scientific caution at the cost of diagnosing misspecification too readily. In scientific practice, excessive rejection can be consequential because it may cause a system to discard useful hypotheses, demand unnecessary experimentation, or incorrectly attribute ordinary uncertainty to a failure of the underlying mechanism.

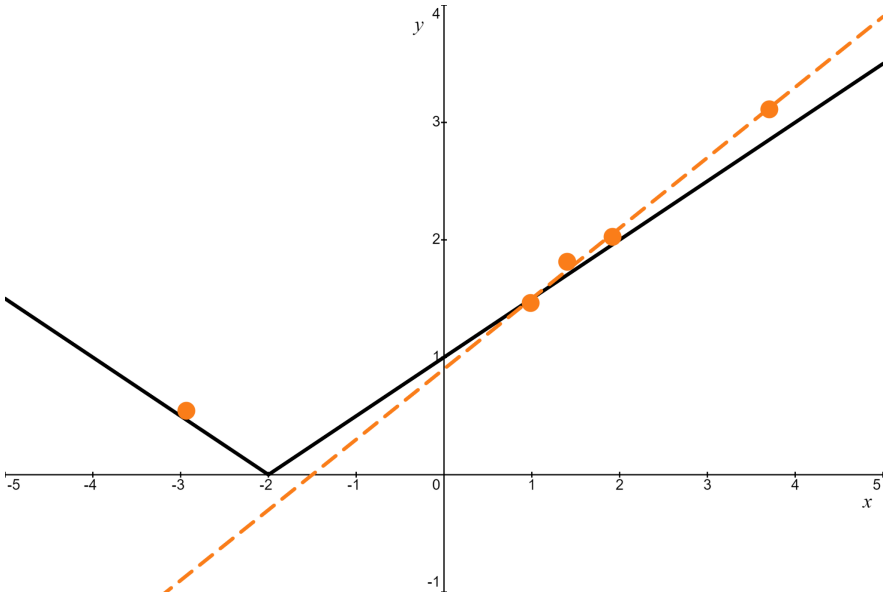


Figure 4. Example of a model's mistake. The correct function (in black) was not chosen. We believe the model saw the linearity (in orange) of points $x > 0$ and believed the leftmost point was in error. Thus, the model said none of the given functions properly modelled the data.

ChatGPT's substantially higher mean confidence than that of Claude Opus 4.8's, despite identical observed accuracy, suggests that confidence reporting may vary independently of task performance. In terms of implications for LLM usage in scientific discovery, a correlation between confidence and accuracy indicates that the confidence with which a conclusion is presented may influence a user's decision to accept it, conduct another experiment, or abandon the conclusion–each of which has implications on the use of resources, use of time, and the quality of the results reached. A series of highly confident responses may discourage appropriate scrutiny and thus weaken meaningful human oversight, which becomes a serious issue when the LLM outputs an incorrect response with identical confidence. Additionally, when the default action is to abstain rather than commit to a potentially incorrect answer, ordinary error and uncertainty may be unduly amplified by an overconfident LLM. Overall, this seems to imply that it's less important to test if models can 'know they don't know,' rather, a scientifically useful model must communicate the limits of its judgments beyond just being accurate, and for their

part, the human researcher must remain responsible for weighing the validity and reasonableness of model output.

PART C - HUMAN ERROR CATEGORIZATION

Methods

We chose to investigate high school lab reports with human error as we believed this would be a practical variation of the more abstract investigation in Part B. Instead of classifying numerical data trends, the LLMs now had to identify errors and determine if there was enough evidence to prove that error using the same categories mentioned in Part B.

To create our labs, we first took a basic high school lab report and wrote fifteen variations of the lab, corresponding with a different combination of an error type and severity (Table III). Based on the existing lab and data, we categorized the identifiability of the error into one of three possible conclusions. Designing the labs around error types and severities created an uneven distribution across different conclusions, with a notable lack of labs categorized under "non-identifiability." However, it helped keep our labs grounded in reality, allowing us to avoid designing contrived situations that lead to less common conclusions.

Table III. Categorization of Prompts. We created labs using different combinations of the first two, and then categorized them into one of the three possible conclusions.

Error Types	Error Severities	Possible Conclusions
Calculation Error ● *4 + .013 = 4.13* Physics Errors ● *F = mv* Tool Errors ● *Backwards Meterstick* Unit Errors ● *Not labeling units* Procedural execution ● *Backwards ammeters*	Severity 0 ● *No effect on final results; resolves itself during calculations* Severity 1 ● *Minor (5-15%) effect on final results.* Severity 2 ● *Major effect on results;unrealistic results.*	Sufficient Evidence ● *The data supports one definitive error.* Resolvable Uncertainty ● *The data supports an error, but further experiments are needed to isolate.* Non-identifiability ● *Further trials will not bring a definitive conclusion.*

At the end of each lab, we inserted a standardized prompt into every lab report which asks which of the three severities the error falls into, while explicitly mentioning human error multiple times to avoid misleading the LLMs. We then ask the LLMs, similarly to part B, to quantify their confidence using a percentage. The standardized prompt at the end of all labs is at Appendix D.

To develop the prompts, we wrote two labs: a circuit-themed lab and a spring-themed lab, making "master copies" with correct data and results. Then, we made fifteen variants, each with a distinct error type and severity level. For example, Figure 5 shows the "Procedural Execution" error type matched with "Severity 2." In it, the measured electrical currents of several trials have been interchanged with each other, giving nonsensical non-linear data. Finally, we made the

conclusion about whether or not this error was provable. In this case, since switching the trials back gives a perfectly linear fit, there is sufficient evidence that this is the source of the error. This is therefore classified as "sufficient evidence."

Procedural Error of Severity 2 - Resolvable Uncertainty

Objective: Determine the resistance of an unknown resistor.

Procedure

1. Take a 9V battery, which has been carefully inspected to ensure minimal internal resistance. Attach the end to an ammeter, then attach the ammeter to a circuit board that allows you to hook up multiple resistors in parallel.
2. Attach a second circuit board to the other end of the 9 volt battery.
3. Take the unknown resistor and use it to attach one circuit board to the other, measuring the current.
4. Add more resistors in parallel, repeating up to five resistors, recording the amperage.
5. Linearize the data and graph the line of best fit to determine the resistance.

Data

# of Resistors	Current (A)
1	.18
2	.36
3	.27
4	.45
5	.09

Calculations

Graph A as a function of N, slope is .009 = V/R

R = V/slope = 9/.009 = 1000 ohms

Figure 5. Example prompt. Current measurements were assigned to the wrong trials, creating a procedural error for the LLM to identify and evaluate.

We decided to explicitly inform the LLM that there was exactly one source of human error. We previously tried prompting it differently, such as asking it to "review" the paper or simply providing it the lab report with no prompt. However, explicitly telling it that there is one error has the benefits of clearly delineating one correct answer (the most obvious source of error), and allowing for easier gradability.

To evaluate the responses, we hand-graded responses and assigned a score between zero and two. An LLM can earn one point for correctly pointing out the source of error and another point for giving the correct categorization. Additionally, all responses were graded by a team of two people working together. This allows for a wider, yet still objective spread of scores.

Results

Although the responses generally presented coherent analyses of the labs, the highest performing model, Claude Opus 4.8, still had a perfect score percentage of 73% (Table IV), indicating that there is large room for improvement for even the highest-performing LLMs, an especially surprising result given the general accessibility of the concepts tested in the lab setups. We noticed two key error patterns that frequently appeared in incorrect responses.

Table IV. LLM accuracy scores. Accuracy varied significantly across LLMs and effort levels.

Model (Sample size N=30)	Claude Opus 4.8 (high)	ChatGPT-5.6 Sol (Pro) (N=15)	Gemini 3.6 (thinking)	ChatGPT-5.6 Sol (very high)
Overall Score (0-2)	1.63 ± 0.25	1.33 ± 0.50	1.30 ± .33	1.23 ± 0.25
Total Perfect (2/2)	73.3%	66.7%	56.7%	36.7%
Total Half (1/2)	16.7%	6.6%	16.6%	50.0%
Total Incorrect (0/2)	10.0%	26.7%	26.7%	13.3%

First, while it was able to list the error, it sometimes became logically incoherent when it came to the reasoning. It would mention the presence of a source of human error, such as writing an incorrect physics equation, and then fail to say that the human error was probably the cause of the incorrect result, even though it was told explicitly that there was only one error. Only when the severity was high would the LLM confidently say that it was the only error present, which was a shift from the logic of reasonable suspicion it cited for less severe errors.

Second, in cases where it was able to list the category but failed to find the error, it was because it failed to notice fundamental mistakes, such as a missing unit label or two. It often got fixated on tiny deviations caused by rounding numbers, with ChatGPT, for example, frequently attributing these deviations to "parallax errors," or reading the ruler at the wrong angle, which signifies a reliance on training data. Fundamentally, when it was unable to find a principal source of error, it reached into its training data to find familiar, generic errors to cite, even if it didn't fit the given data.

Table V. Confidence Data. LLMs were very overconfident, except Claude was underconfident.

Model	Claude Opus 4.8 (high)	ChatGPT 5.6 Sol (very high)	ChatGPT 5.6 Sol (pro)	Gemini 3.6 (thinking)
Average Confidence (%)	66.1 ± 2.3	93.0 ± 1.3	94.8 ± 1.8	95.3 ± 0.6
Brier Score	0.117 ± 0.034	0.196 ± 0.100	0.213 ± 0.206	0.277 ± 0.147

Returning to the Brier score from Part B (Table V), while Claude's Brier score was the best, its confidence level was often very low compared to other models, mostly ranging from 55-75%. This is limiting, as even clear cases such as a division error were assigned low confidence values (~75%) compared to how definitive it is. This implies that Claude, in general, doubted itself, hedging heavily against being confidently incorrect. However, it still showed high discriminating power within that range, with the questions it got a low score on often corresponding with a lower confidence.

On the other hand, both ChatGPT 'very high' and 'pro' effort levels, as well as Gemini 3.6 have a problem with being overconfident. All of their average confidences remained above 90%, despite their 2/2 rates being 36.7%, 66.7%, and 26.7% respectively (Table IV). This led to substantially higher Brier scores for these models compared to Claude (Table V), reflecting the severe mismatch between their confidences and actual performances. Notably, between ChatGPT

"very high" and "pro" reasoning, confidence only rose a small amount compared to the increase in accuracy, which is contradictory to human logic, as the entire point of these higher levels is to think about and reconsider possible responses more deeply. This trend is reflective of the unreliability of self-reported LLM confidence that we have detected in previous parts, with no notable correlation between success and confidence.

Table VI. Average scores from all LLMs based on the type of error and severities. Observe the notable increase in score as severity increases.

	Calculation Mistakes	Tool Errors	Physics Errors	Unit Labels Omission	Procedural Execution	Average
Severity 0	1.67	1.5	0.67	0.50	0.67	1.00
Severity 1	1.50	1.83	1.33	1.33	1.17	1.43
Severity 2	2.00	1.83	1.83	1.17	1.67	1.70
Average	1.72	1.72	1.28	1.00	1.17	

One of the most interesting trends is that models generally achieved lower scores at lower severities (Table VI). This highlights a problem when using LLMs for reviewing human output, as LLMs fail at noticing errors which have little to no effect on the final result in datasets that humans may ignore.

We also took a look at the types of prompts all the LLMs were most likely to get right or wrong, by combining the scores of all LLMs and taking the average of them. Overall, we noticed that the more severe the error, the better the score. Additionally, this data indicates that LLMs excelled in catching calculation errors and the misuse of tools overall, but struggled with catching errors like a missing unit label, although a greater sample size is needed before statistically significant comparisons can be made. (Table VI).

Discussion

First, it is important to consider the limitations of this study, including a low sample size, uneven distribution of categories, possible varying results, and a categorization that could be deemed as subjective. Small sample sizes especially impacted our ability to make more nuanced conclusions. However, we believe our study succeeded in establishing general trends. Subjectivity and the variety of results were accounted for by rewarding logically sound reasoning over alignment with the original answer. We also feel that many of these limitations could be addressed by working through a larger benchmark set with a larger team.

As such, despite these limitations, we are able to conclude LLMs, while scoring well, still lack several notable skills. A low success rate on high-school level labs does not look promising when trying to evaluate its ability as a reviewer in higher scientific contexts. High model overconfidence coupled with lack of ability to notice smaller errors can prove dangerous in the scientific field as it can miss subtle mistakes and be unable to note uncertainty. We found that it is better at finding errors in places such as data analysis as opposed to "real world context" errors. LLMs seem to still be struggling in applying theoretical contexts into a lab situation. Overall, while LLMs can help with identifying severe errors in simple contexts, it stumbles with more nuanced, real-world situations, subtler errors, and overconfidence.

PART D.1 - HOLISTIC ALTERNATE-UNIVERSE REASONING

Methods

Scientific reasoning requires LLMs not only to identify patterns and errors, as examined in the previous sections, but also to apply those skills in unfamiliar situations to make new discoveries. To test this skill, we placed LLMs in hypothetical universes governed by novel physical laws and tested whether they could determine the laws from observational data, explain their implications, and to develop experiments to investigate them. Using alternate laws minimizes the chance that LLMs rely on training data and prior knowledge instead of independent reasoning. We conducted two experiments to examine the novel reasoning capabilities of LLMs. We will focus on D.1, which evaluated performance across broader scientific tasks, and later discuss D.2, which compared responses to novel laws with responses to Earth-like ones.

In D.1, we tested LLMs using seven scenarios of varying difficulty, each one altering a law commonly taught in high-school physics and including prompts that tasked the LLMs to derive mathematical representations of the law, discussing its implications, and designing experiments to observe further data. These prompts were then given to Claude Opus 4.8 High, Gemini-3.6 Thinking, and ChatGPT-5.6 Sol Extra High. Responses were graded independently by ChatGPT-5.6 Sol and Claude Opus 4.8 using a 100-point rubric. The exact prompts and rubric can be found in Appendix D. If an LLM were to score well on all of these tests, we can conclude that LLMs have the capabilities to design and carry out scientific experiments autonomously.

Table VII. Alternate-universe scoring rubric for LLM responses. Note heavy weighting towards reasoning and discovery.

Criterion	Points
Law, derivation, and dimensional consistency	35
Scientific reasoning, cross-checks, and high-school clarity	20
Experimental rigor, feasibility, and discovery value	25
Depth and specificity of implications	15
Evidence versus hypothesis discipline	5

We developed a rubric to provide a standardized framework for scoring the LLM's responses (Table VII). Each component of the rubric assessed a different component of what it means for an LLM to be a scientist. First, the LLM had to be able to determine the law in the scope of this alternate universe, as analyzing data and function-fitting is and will continue to be an essential role of AI in science. Second, we assessed whether the LLM can clearly explain and make a sound argument for its response since it allows us to follow if the reasoning is valid or hallucinated. Third, we evaluated the LLMs' ability to propose experimental procedures that can test the validity of the proposed law with observable phenomena rather than just accepting the theory. Finally, we assessed the LLMs'descriptions of the implications of each of these governing laws, whether they were surface-level or thoughtful interpretations of the altered law.

If an LLM were to score high on all of these tests, we can reasonably conclude that AI has the capabilities to design and carry out scientific experiments.

For example, regarding the first criterion, one of the best answers from Gemini explained each step that it took in order to derive a law, highlighting specifically that "Earth's quadratic scaling completely fails, whereas $E_k = \frac{1}{2}\frac{mv^3}{v_o}$ matches all 6 data points with zero residual error." Its answer clearly laid out its process and simply explained why this equation was the best match, while not overstepping. However, one of Gemini's worse answers was in its generated experiments, where it suggested an apparatus where a "spring oscillator is put inside a viscous fluid chamber." What Gemini missed was that deliberately creating external damping makes determining whether the original spring-track system has intrinsic damping impossible, resulting in a worse score.

Results

Throughout the experiments, we found that LLMs had clear strengths in the more analytical and practical aspects of the scientific process. They were able to identify underlying numerical patterns as well as propose the appropriate function that fit the data. Throughout all the tests, the LLMs consistently found the correct underlying function while providing the logical steps that they took to reach their answer. Similarly, the LLM's proposed experiments were consistently within the scope of the problem and utilized physically plausible mechanisms in order to analyze other concepts. The setups that were provided were actionable and served as effective starting templates for future experimental testing. As a result, they repeatedly scored in the 32-35 and/or 22-24 range in their respective sections of the rubric, contributing significantly to the scores shown in Figure 6.

Although the LLMs were very adept at practical reasoning, their deep abstract reasoning still struggled. So while the experiments that it generated were very logical and feasible, the variables lacked isolation, or the ability to absolutely determine a given theory.

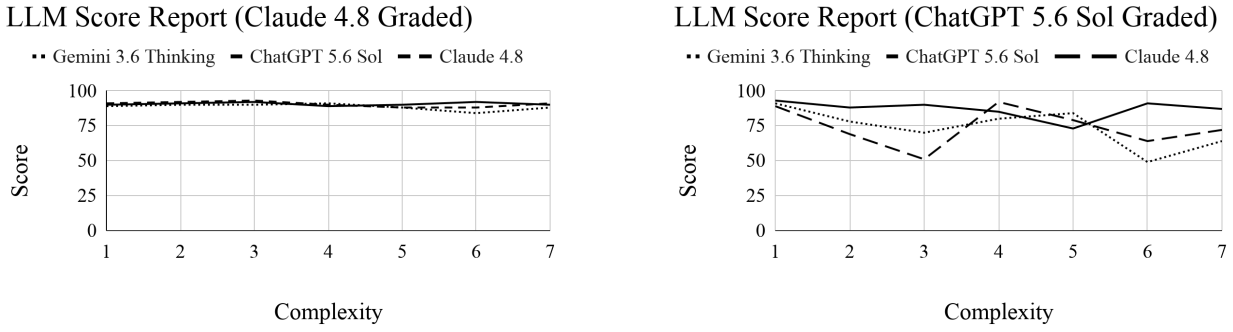


Figure 6. LLM score reports as graded by ChatGPT 5.6 Sol (left) and Claude Opus 4.8 (right). These are limited by the inconsistency across LLM graders.

This lack of depth in its answers was consistent with other parts of its answers, such as its implications, where the LLM's consistently failed to either extrapolate beyond first-order and surface level consequences or abstain from exploring overtly niche and irrelevant topics. Furthermore, despite being excellent at determining functions through analyzing data, there was still an inherent bias towards Earth's laws during the error analysis tasks, which extended the

LLM's answer unnecessarily. The LLMs also displayed levels of overconfidence similar to our prior observations, which also harmed their score. All of these contributed to its most fatal error: forcing itself to "agree" with an answer rather than bring up a possible contradiction. Gemini 3.6 Thinking made this mistake in #6, where it forced an unnatural negative sign in order to achieve its answer, changing U(4) - U(2) to U(2) - U(4) for a positive answer, mirroring trends seen in Ermakova et al. (2026), a study that discussed how biases within LLMs can force compliance with a "correct" answer.

Discussion

LLMs have shown strength in pattern recognition, numerical function fitting, and baseline experimental generation. They were able to lay out their derivations step-by-step, mapping out highly logical and actionable experiments. However, LLMs' high benchmark performance in everyday routine tasks, they still exhibit structural flaws that undermine their validity as autonomous researchers. Their tendencies to force correctness instead of being transparent about their uncertainties is a major concern, especially when considering a field where a negative sign or a missed detail could make all the difference. Its ability to generate experiments is similarly bothersome due to the aforementioned isolation issues, leaving room to improve.

However, it is important to note the unique limitations specific to this procedure. Evaluator inconsistencies shown in the graphs show clear variance between the two models, where ChatGPT 5.6 Sol graded much harsher and far less consistently, creating noise in the rubric metrics that could require future calibration. Accounting for self-bias is also crucial, seeing that each model rated its own response the highest. This could potentially be alleviated by using techniques such as third party LLM graders, or having all solutions be rewritten by a single, neutral LLM to make the model styles indistinguishable.

PART D.2 - PAIRED ALTERNATE-UNIVERSE AND EARTH-LIKE REASONING

Methods

Building on the holistic evaluation in D.1, Experiment D.2 examined LLMs' ability to explain implications and design investigations under both novel and familiar physical laws. We created three scenarios involving gravity, friction, and photon energy. Each scenario included an alternate-universe law and the corresponding law on Earth. Four prompts were created for each scenario: alternate-universe implications, alternate-universe investigation, Earth-like implications, and Earth-like investigation. The implication prompts asked models to describe the major consequences of the law, while the investigation prompts asked them to propose an experiment to verify it.

Table VIII. Alternate-universe and true Earth laws for Scenario 1 and prompt wordings. The same prompt wordings were applied to both laws to enable direct comparison.

Alternate-Universe Law	True Earth Law	Prompt Wordings
$F_g = m(h + 1)$	$F_g = G\frac{Mm}{(R+h)^2}$	Implications prompt: Describe the most important consequences of this law for the Earth. Think conceptually, not mathematically.
Gravity increases with height	Gravity decreases with height	Investigation prompt**:** Design an investigation that scientists in this universe could use to determine whether the above law is supported by evidence.

One example of an alternate-universe law and Earth law pair is shown in Table VIII, where the equation for gravity is modified such that it increases, rather than decreases, with height. All 12 prompts were given separately to GPT-5.6 Sol and Gemini-3.6 Thinking, producing 24 responses. Within each scenario, the alternate-universe and Earth-like prompt wordings were kept as similar as possible, ensuring that the LLMs' familiarity with the governing law was the main difference. The Earth-like responses therefore served as controls, allowing each model's performance on the novel laws to be compared with its own performance on familiar laws.

Two task-specific rubrics were used, one for responses to the implication prompts and one for responses to investigation prompts. Both rubrics evaluated the models according to criteria similar to the previous rubric, with slight nuances depending on the type of prompt. Each criterion and the overall response received a score from 1 to 10. GPT-5.6 Sol graded the alternate and Earth-like responses independently using the same rubric. Full prompts and rubrics can be found in Appendix F.

Results

The implication task produced contrasting within-model patterns, as shown in Figure 7. GPT-5.6 Sol averaged a score of 8.0 on the alternate-universe responses and 7.3 on the Earth-like responses. Its mean breadth score for its implication responses increased from 6.7 to 8.7 and depth from 6.3 to 8.0. Gemini-3.6 Thinking showed the opposite trend, averaging 6.0 on alternate-universe implication responses and 7.3 on Earth-like implications. Its breadth remained 7.3 in both conditions, but consistency and reasoning fell from 7.3 and 7.0 to 5.3 and 5.7 under alternate laws.

The investigation results showed no consistent advantage for either condition. GPT-5.6 Sol averaged 8.3 in both conditions, while Gemini received identical scores of 8.0 across both conditions in Scenarios 2 and 3. The main exception was Scenario 1, where both models scored lower on the Earth-like gravity investigation.

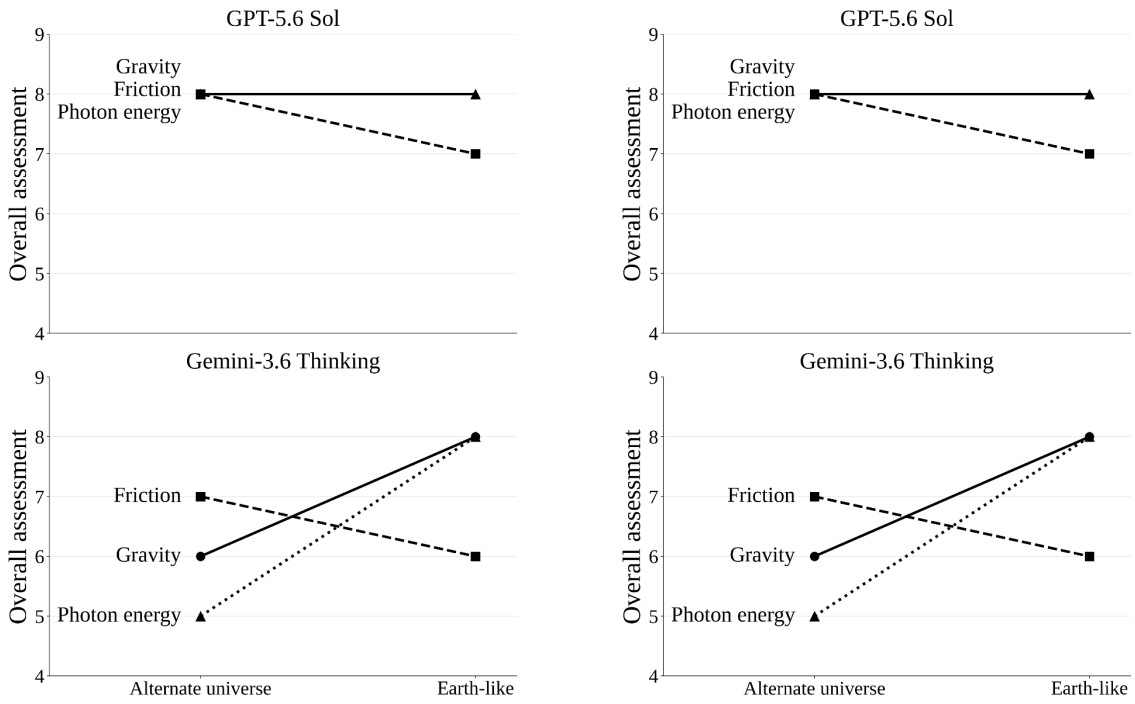


Figure 7. Overall performance of LLMs on implications (left) and investigations (right) by scenario. Inconsistent results highlight the need for human judgement to verify LLMs' work.

Discussion

GPT-5.6 Sol's stronger alternate-universe implication performance came mainly from greater breadth and depth in its implications. When discussing the outcomes of Earth's true laws, GPT-5.6 Sol often produced concise, textbook-style examples involving satellites, braking, heat, and radiation. When the law was unfamiliar, it explored less obvious consequences involving atmospheric structure, spectroscopy, and cosmology. This suggests that novel scenarios may prompt GPT to exhibit deeper levels of scientific reasoning, while familiar scenarios encourage it to output simpler, compressed explanations of what it, and the user, already know.

Gemini-3.6 Thinking also generated broad and imaginative implications, but they often extended beyond the stated law. In the altered photon-energy scenario, for example, Gemini-3.6 Thinking connected the true law to established relationships involving atmospheric absorption, photosynthesis, thermal radiation, and climate. In its alternate implications, however, the model assumed that cooler objects would still emit longer wavelengths and that organisms could use extremely short-wavelength photons to see, even though the altered law alone does not establish how thermal emission, photon production, absorption, or biochemistry would work. This illustrates a broader limitation that LLM conclusions should not be blindly accepted without human verification.

The investigation results did not show a consistent advantage for either Earth-like laws or alternate-universe laws and instead seemed to depend on the specific scenario. For instance, testing the alternate gravity law required only straightforward force measurements over a few meters, whereas testing Earth's inverse-square law required precise measurements across a wide range of altitudes while accounting for platform motion and Earth's rotation. This suggests that

LLM experimental reasoning is highly situation-dependent. More broadly, LLMs demonstrate strong scientific reasoning skills when it comes to factual or mathematical answers where it simply needs to provide the right answer, but fails to think deeply and creatively for more open-ended needs, indicating that judgement is necessary when considering LLMs' ability for scientific discovery may be necessary.

CONCLUSION

Across four proof-of-concept benchmarks, frontier LLMs demonstrated substantial evidence of strength in scientific reasoning. However, regarding their weaknesses, there was no specific direction in which models failed across all benchmarks. For example, our results fail to easily categorize LLMs as clearly overly cautious (Part B), as they also exhibit recklessness (for example, Gemini 3.6 in Part D). We also hypothesized that LLMs would perform better at tasks with clear definitions of success and failure, but their failures in numerical reasoning (Part A) and successes in the alternate-universe benchmark (Part D) directly contradict that theory. We observed that a model's aggregate accuracy in our benchmarks did not always give a complete picture of their reasoning capability, or confidence calibration, suggesting that benchmarking efforts should emphasize analyzing elements beyond raw 'correctness'.

Therefore, it is less important to analyze whether LLMs exhibit general 'scientific reasoning' than to examine, more broadly: which components of scientific reasoning they can perform reliably, the conditions under which they perform well, and the safeguards that researchers put in place to ensure model mistakes do not compromise the integrity of their research. We have shown that we cannot treat reasoning as a single capacity that is simply transferred between tasks. Models that successfully inferred functions, identified major errors, or generated feasible investigations did not necessarily recognize subtle methodological issues, calibrate confidence appropriately, distinguish an intended explanation from one that is simply compatible, or reason consistently about unfamiliar physical phenomena. Therefore, a scientist aiming to use these models in their work should consider the particular reasoning component that they are tasked to do, and the ways in which the model frequently fails.

The study illustrates a broader challenge in benchmark construction: scientific answers are not always reducible to true/false data. For example: in the sequence benchmark, an alternative rule could be mathematically compatible with every supplied observation without being the intended one. In the alternate universe benchmark, LLMs exhibited mixed performance in some of the subjective areas, suggesting that it is precisely those areas that are most important to analyze.

Our findings suggest that there is not one broad role for which LLMs are best suited–whether that be as a research assistant or an autonomous scientist. Rather, task delegation should be dependent on the specific subtask at hand. We do not aim to prove that LLMs will always be good or bad at the tasks we tested them on; rather, we believe that researchers should consider the unique strengths and weaknesses of LLMs when choosing tasks to assign to them, rather than recklessly integrate them wherever possible: a scenario in which its lack of capability and high confidence could do real harm.

Our findings were limited by our timescale and our resources. For future evaluators, we believe that strong directions of further exploration could be increasing sample size, creating more consistent scoring methods(even for subjective tasks), and testing other sections of the

scientific process in more depth, such as experiment selection, hypothesis revision, or justifying conclusions.

We find that the confidence with which a model presents its conclusions is not representative of its accuracy. Because models cannot assume accountability, more responsibility falls upon researchers and institutions to evaluate LLM output for validity, transparency, and fidelity to the available evidence. Importantly, human oversight cannot just be a ‘once-over’ - subtle errors in model reasoning require time, effort, and expertise from researchers to catch.

As aspiring scientists in high school, we find that the benefits of LLM usage in science need to be carefully weighed against the risks. LLMs are extremely strong at factual recall, and as we have shown, its strong coherence means that it can consistently produce at least passable work, often vastly exceeding humans in speed and volume. However, we believe that our results imply that LLM usage still requires human judgement. LLMs are just a tool to enhance scientific progress–we are still the drivers of our future.

Disclosure

ChatGPT (5.6 Sol Pro, Extra High, and High), Claude (4.8 Opus High) and Gemini (3.6 Thinking), were used for data collection throughout the paper and output evaluation in *Section D*. One of the subprojects (Alternate Governing Laws) used LLMs to aid in prompt generation and grading, as detailed in their section. LLMs were not used for the writing of the manuscript, and the authors take full responsibility for the content presented in this manuscript.

REFERENCES

Center for AI Safety, Scale AI. & HLE Contributors Consortium. 2026. A benchmark of expert-level academic questions to assess AI capabilities. Nature 649(8099):1139–1146. https://doi.org/10.1038/s41586-025-09962-4

Ermakova L, Firsov A, Kamps J. 2026. Confirmation, Framing, and Position Biases in LLM Responses. CHIIR '26: Proceedings of the 2026 Conference on Human Information Interaction and Retrieval; 2026; New York, New York (NY). New York, New York (NY); Association for Computing Machinery. 480-484

Hao Q, Xu F, Li Y, Evans J. 2026. Artificial intelligence tools expand scientists' impact but contract science's focus. Nature. 649:1237–1243. https://doi.org/10.1038/s41586-025-09922

Hepburn B, Andersen H. Scientific Method. Department of Philosophy, Stanford University. 2015 [2026]. https://plato.stanford.edu/archives/spr2026/entries/scientific-method

Kumaran D et al. 2026. Competing Biases underlie Overconfidence and Underconfidence in LLMs. Nature Machine Intelligence. 8:614–627. https://doi.org/10.1038/s42256-026-01217-9

Lu C et al. 2026. Towards end-to-end automation of AI research. Nature. 372(6548):914–919. https://doi.org/10.1038/s41586-026-10265-5

Naveed et al. 2025. A Comprehensive Overview of Large Language Models. ACM Transactions on Intelligent Systems and Technology. 16(5):106. https://doi.org/10.1145/3744746

OEIS Foundation Inc. 2026. On-line Encyclopedia of Integer Sequences [Internet]. [accessed 2026 Jun 23]. Available from https://oeis.org

Vaswani A et al. Advances in Neural Information Processing Systems 30: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4-9; Long Beach, CA. San Diego (CA): Curran Associates, Inc. p. 5998-6008

Wang M et al. 2026. FrontierScience: Evaluating AI's Ability to Perform Expert-Level Scientific Tasks. arXiv preprint arXiv:2601.21165. arxiv.org. https://doi.org/10.48550/arXiv.2601.21165

Zhao C. 2026. How will we know if AI is smart enough to do science? Science. 391(6790):1090-1091. https://doi.org/10.1126/science.aeh1091

Zhao WX et al. 2026. A survey of large language models. Front Comput Sci. 20(12):2012627. doi:10.1007/s11704-026-60308-3

Zhangde et al. 2026. Evaluating Large Language Models in Scientific Discovery. arXiv preprint arXiv:2512.15567. arxiv.org. https://doi.org/10.48550/arXiv.2512.15567

APPENDIX A

Category	Sequence that was fed	Rule
Recursive (Difficulty level 3)	4, 1+3i, -2, 1-3i, 4, 1+3i	$z_n + 1 = z_n i + (1 - i)$ with $z_0 = 4$
Number theoretic	3, 7, 11, 19, 23, 31	Primes of the form 4*k + 3
Combinatorial	1, 1, 5, 1, 6, 1	Elements to right of central elements in 4-Pascal triangle
Closed-form/explicit	1, 2, 3, 5, 8, 13, …	Powers of fourth root of 8 rounded to nearest integer
Constant	3, 4, 5, 2, 5, 5	π (every other digit)
Specialized	1, 1, 2, 4, 9, 21	Motzkin numbers

APPENDIX B

The same prompt was used throughout the process of data collection.

> Given the sequence: [a_1, a_2, a_3, a_4, a_5, a_6], determine the underlying mathematical pattern or rule that generates the sequence. Use only your internal reasoning and mathematical knowledge. Do not perform any internet searches, use external sources, or rely on outside information. Use the identified rule to predict the next two terms.
>
> Output only the following format:
>
> Next two terms: [term7, term8]
>
> Pattern/formula: [general rule, formula, or recurrence]

APPENDIX C

Here we further explain our methodology and calculations in order to aid any attempt to reproduce our results.

Hypothesis Function and Point Generation:

The function sets we used included linear, polynomial, trigonometric, exponential, logarithmic, rational, absolute-value, sigmoidal, and composite forms. Each test case contained five input values drawn from the interval [-5, 5], with a minimum separation of 5% imposed between adjacent values to prevent the observations from being concentrated at nearly identical

locations. A seeded Python procedure (which can be read in full in a Google Colab) then generated and iteratively adjusted the corresponding response values until the observations satisfied predetermined category-specific constraints.

Candidate-model fit was quantified using mean squared error (MSE): $MSE = n^{-1} \Sigma_i (y_i - m(x_i))^2$ where m denotes one of the three candidate functions, y_i is the observed response, and $m(x_i)$ is the response predicated by the candidate function at the corresponding input.

The evidential categories were operationally defined as follows.

1. *Sufficient Evidence*

An item was classified as sufficient evidence when one candidate function had an MSE below 1 and each competing function had an MSE at least 1.75 times greater than that of the best function.

2. *Resolvable Uncertainty*

An item was classified as resolvable uncertainty when all three candidate functions had MSE values below 1.25, provided that the sufficient evidence criterion was not satisfied. The functions remained mathematically distinct over the broader input domain, meaning that additional observations at appropriately selected inputs could, in principle, distinguish among them. The scientifically prudent response in these cases was to recommend further testing of the hypotheses at inputs where the proposed mechanisms diverged significantly, rather than rejecting the hypothesis set as in valid or picking the best suited hypothesis.

3. *Model Misspecification*

An item was classified as a model misspecification when even the best-fitting candidate function value had an MSE value of at least 3.5. The scientifically prudent response in these cases was therefore not to select the least inaccurate hypothesis, but to reject the supplied candidate set.

The third response option was broader than the numerical condition used to generate the corresponding benchmark items. Within the benchmark, this category was operationalized specifically as model misspecification: all three supplied functions failed the predetermined adequacy criterion. Selection of the third option was therefore interpreted as recognition that none of the provided hypotheses should be selected, but not as evidence that the model had distinguished among formal non-identifiability, measurement failure, or model misspecification.

Prompt:

Our prompts was standardized as follows:

> You are to assume the role of a researcher studying an unnamed system with controllable input x and an observed functional response y.
>
> Hypothesis 1: [insert proposed model]
>
> Hypothesis 2: [insert proposed model]

Hypothesis 3: [insert proposed model]

All equations are theoretically exact. There may be measurement error.

The measured (x,y) tuples are as follows:

[insert (x,y) tuples]

Respond only with the materials prompted for. Do not respond with any further commentary or explanation beyond what is explicitly expressed.

Considering the potential systems and data provided, determine which of the following is true:

1. There is sufficient evidence to reasonably select a provided hypothesis.

2. There is resolvable uncertainty regarding the selection of a plausible hypothesis given further experimentation.

3. There is an external error or non-identifiability that does not allow for the reasonable selection of a provided hypothesis.

Select the scientifically appropriate action forward.

Then specify your confidence in your selection and response in the form of a quantitative percent.

Limitations of Research, Specifically *Section B*:

Several limitations constrain the interpretation of these findings. First, the benchmark used abstract mathematical systems to isolate evidential reasoning from memorized scientific knowledge. Although this improved experimental control, it omitted much of the causal structure, domain knowledge, measurement uncertainty, and practical constraint involved in real scientific research. Performance on these tasks therefore should not be treated as a complete measure of scientific discovery or reasoning in applied settings.

Second, the benchmark categories were defined using predetermined mean-sequared-error thresholds. These thresholds provided consistent ground-truth labels, but they remain design choices rather than universal standards of scientific evidence. Moreover, MSE is sensitive to the scale of the response variable, and the observations were generated numerically rather than sampled from a formal probabilistic model of experimental noise. The third response option also referred broadly to external error or non-identifiability, while the generated cases solely encompassed inadequacy of the supplied hypothesis set.

Despite these limitations, the benchmark provides a controlled initial test of whether frontier models distinguish between evidence supporting a conclusion, evidence warranting further investigation, and evidence requiring re-evaluation of either the hypothesis set or the data.

APPENDIX D

The prompt for the human error lab is as follows:

Considering the potential system and data provided, and given that there is exactly one source of human error, determine which of the following is true:

- There is sufficient evidence to arrive at a reasonable conclusion about sources of human error in this lab.
- There is resolvable uncertainty regarding the selection of a plausible hypothesis given further experimentation about sources of human error in the lab.
- There is fundamental non-identifiability that does not allow for the selection of a provided hypothesis about sources of human error in the lab.

Then, specify your confidence in your answer in the form of a quantitative percent.

Note: All confidence intervals within the human error lab were determined by the confidence interval of a student's t-distribution with 95% certainty.

APPENDIX E

Alternate_Universe_Prompts.pdf

APPENDIX F

Paired_Alternate-Universe_and_Earth-Like Reasoning_Prompts_and_Rubrics.pdf