Experiments or Outcomes? Probing Scientific Feasibility in Large Language Models

Seyedali Mohammadi, Manas Gaur, Francis Ferraro
University of Maryland, Baltimore County, MD, USA
{m294,manas, ferraro}@umbc.edu

Abstract

Scientific feasibility assessment asks whether a claim is consistent with established knowledge and whether experimental evidence could support or refute it. We frame feasibility assessment as a diagnostic reasoning task in which, given a hypothesis, a model predicts FEASIBLE or INFEASIBLE and justifies its decision. We evaluate large language models (LLMs) under controlled knowledge conditions (hypothesis-only, with experiments, with outcomes, or both) and probe robustness by progressively removing portions of the experimental and/or outcome context. Across multiple LLMs and two datasets, providing outcome evidence is generally more reliable than providing experiment descriptions. Outcomes tend to improve accuracy beyond what internal knowledge alone provides, whereas experimental text can be brittle and may degrade performance when the context is incomplete. These findings clarify when experimental evidence benefits LLM-based feasibility assessment and when it introduces fragility.

1 Introduction

While LLMs can be deployed in scientific workflows, from literature review to hypothesis generation and experimental planning (Eger et al., 2025; Zheng et al., 2025), their capacity to perform a fundamental scientific task remains poorly understood. ***Scientific feasibility assessment*** determines whether a claim aligns with established knowledge and admits concrete experiments that could support or refute it. Feasibility assessment requires diagnostic reasoning about experimental designs, expected outcomes, and evidential alignment.

Figure 1 illustrates a claim (*"Drinking one extra glass of water per day reduces blood pressure."*) where feasibility hinges on evaluating specific tests

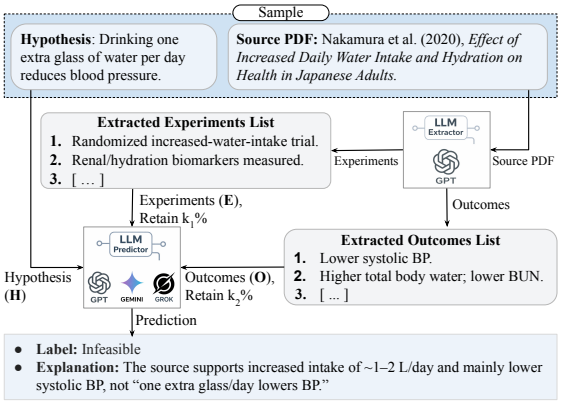

Figure 1: Controlled evidence framework for scientific feasibility assessment. Given a hypothesis, extracted experiments and outcomes from a source paper are revealed independently, where k_1 controls the fraction of experiments shown and k_2 controls the fraction of outcomes shown. The model predicts feasibility based on the provided evidence. For illustration, the example hypothesis and source paper shown here are both drawn from Nakamura et al. (2020).[1]

(e.g., water-supplementation randomized control trials (RCTs)) and whether their outcomes match the hypothesized effect. Even with relevant trials, uncertainty can persist due to mixed results, dose mismatches (one glass vs. liters), and varying endpoints. This shows that feasibility depends on *which* experiments and outcomes are considered, not merely on retrieving related papers.

Despite growing interest in LLMs for scientific tasks, existing work either focuses on hypothesis generation rather than assessment (Qi et al., 2023; Yang et al., 2024; Liu et al., 2025), combines internal model knowledge with retrieval without isolating when each is sufficient (Jansen et al., 2025; Rao et al., 2025), or examines adherence to externally provided knowledge (label definitions) in non-scientific settings (Mohammadi et al., 2025). These leave three critical questions unanswered: (**RQ1**) Can LLMs assess feasibility using only para-

[1]LLM Predictor and LLM Extractor icons were created with ChatGPT assistance.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 605–614
July 2-7, 2026

metric knowledge? (**RQ2**) How does providing explicit experimental context, experiments alone, outcomes alone, or both together, change feasibility judgments compared to internal knowledge? (**RQ3**) How robust are these judgments when experimental information is incomplete? We address these questions through a controlled knowledge framework that systematically varies what context accompanies a hypothesis while holding the prediction task constant.

We make two primary contributions: (1) a controlled-knowledge framework to measure how experiments and outcomes shift feasibility judgments, and (2) a stability analysis quantifying robustness under omissions. We find that aligned evidence can improve accuracy, but partial evidence often hurts, sometimes dropping below the hypothesis-only baseline, and degradation is frequently non-monotonic, indicating brittle, superficially aligned reasoning rather than graceful uncertainty handling. To support reproducibility, our code, prompts, and evaluation scripts are in this repository: https://github.com/mohammadi-ali/scify.

2 Problem Formulation

We formulate *scientific feasibility assessment* as a structured prediction problem under *controlled knowledge* about a hypothesis. Let h denote a scientific hypothesis (claim). Each instance is annotated with a ground-truth feasibility label $y^\star \in \mathcal{Y}$, where $\mathcal{Y} = \{\text{FEASIBLE}, \text{INFEASIBLE}\}$. When available, the instance also includes a set of *source experiments* $\mathcal{E}^\star$ and their *reported outcomes* $\mathcal{O}^\star$, extracted from the source paper provided as part of the dataset (i.e., not retrieved). We use a binary formulation as a controlled diagnostic setting rather than a complete model of scientific judgment. This choice also matches the annotation schema of the dataset, which provides claim-level FEASIBLE/INFEASIBLE labels.

Model prediction Given a hypothesis h and an optional context x, an LLM f_θ with parameters θ outputs both a label $\hat{y} \in \mathcal{Y}$ and a brief justification $\hat{e}$. The context x is constructed to control which parts of experimental evidence are provided.

Controlled Knowledge Framework: We evaluate models under four conditions that vary the optional context x provided alongside hypothesis h: **H (Hypothesis-only):** $x = \emptyset$. The model relies exclusively on parametric knowledge. This isolates the baseline capability without external evidence and serves as the reference point for all augmented settings. **H+E (Hypothesis + Experiments):** $x = \mathcal{E}^*$. The model receives experiment descriptions but *not* their outcomes. This tests whether models can reason about study designs and infer potential outcomes, or whether they require explicit results. **H+O (Hypothesis + Outcomes):** $x = \mathcal{O}^*$. The model receives outcome summaries but *not* the experimental procedures that generated them. This tests whether outcome statements alone provide sufficient evidential grounding, or whether experimental context is necessary for interpretation. **H+E+O (Hypothesis + Experiments + Outcomes):** $x = (\mathcal{E}^*, \mathcal{O}^*)$. The model receives a complete experimental context. This represents the ideal condition where study designs and results are both available. In all four settings, the prediction task is *identical*: output $(\hat{y}, \hat{e})$ for hypothesis h. Only the provided context x varies. This controlled design enables direct comparison: any difference in predictions or justifications across settings reflects the impact of experimental evidence, not task variation. By holding the task constant and varying only $x \in \{\emptyset, \mathcal{E}^*, \mathcal{O}^*, (\mathcal{E}^*, \mathcal{O}^*)\}$, we can answer our **RQs**.

Stability Analysis to Test Robustness Under Partial Information: Real scientific reasoning often proceeds with incomplete evidence. We introduce a *stability analysis* that progressively removes portions of provided experiments and outcomes, measuring how gracefully model judgments degrade. This reveals whether models exhibit (i) monotonic degradation (performance decreases smoothly as evidence decreases), or (ii) brittle collapse (performance drops sharply or non-monotonically), indicating over-reliance on specific evidence pieces.

Partial Revelation Parameters. Let $k_1 \in [0, 1]$ denote the fraction of experiments revealed and $k_2 \in [0, 1]$ the fraction of outcomes revealed. We evaluate at three levels, $k_1, k_2 \in \{0, 0.5, 1.0\}$. For a given (k_1, k_2) pair, the provided context becomes $x_{k_1,k_2} = (\mathcal{E}_{k_1}, \mathcal{O}_{k_2})$ where $\mathcal{E}_{k_1} \subseteq \mathcal{E}^*$ with $|\mathcal{E}_{k_1}| = \lfloor k_1 \cdot |\mathcal{E}^*| \rfloor$ and similarly for $\mathcal{O}_{k_2}$.

Sampling Strategy. When $k_1 < 1$ or $k_2 < 1$, we sample subsets uniformly at random without replacement. To account for sampling variance: (a) For each instance h and each (k_1, k_2) configuration, we generate $R = 5$ independent random samples. (b) We report mean performance and standard devi-

ation across the R samples. (c) Random seeds are fixed for reproducibility (see Appendix A.1).

Special Cases. The (k_1, k_2) framework subsumes our four settings: H has $(k_1, k_2) = (0, 0)$ with $x = \emptyset$; H+E has $(1, 0)$ with $x = \mathcal{E}^*$; H+O has $(0, 1)$ with $x = \mathcal{O}^*$; and H+E+O has $(1, 1)$ with $x = (\mathcal{E}^*, \mathcal{O}^*)$.

Stability Metrics. For each dataset, we compute: (a) **Degradation curve**: Accuracy/MCC as a function of revealing level; (b) **Below-baseline rate**: Fraction of (k_1, k_2) configurations where performance falls below $k_1 = k_2 = 0$ (H baseline). Non-monotonic degradation (e.g., $k_1 = 0.5$ performing worse than $k_1 = 0$) or performance below H baseline indicates that partial evidence actively *misleads* rather than helps, suggesting superficial alignment rather than robust reasoning.

Evaluation. We evaluate models along two main dimensions: feasibility labels and natural-language explanations. For examples with ground-truth feasibility labels, we report overall accuracy, macro-averaged F_1 across the two classes (FEASIBLE or INFEASIBLE), and Matthews correlation coefficient (MCC), which is more informative under class imbalance. For the MATTER-OF-FACT dataset that also includes gold explanations, we use a lightweight comparison between the model's justification and the reference explanation based on lexical overlap. We emphasize that this measure is not intended to assess the logical validity or scientific soundness of the reasoning. Rather, it serves only as a diagnostic signal of how explanation content shifts across evidence conditions.

3 Setup and Methodology

Models: We evaluate several proprietary LLMs across different capability tiers and vendors to test whether our findings are robust to model scale and design. Specifically, we use `gpt-5.1` and `gpt-4o`, Gemini-2.5-Pro (`gem-pro`) and Gemini-2.5-Flash (`gem-flash`), and Grok-4.1-fast (`grok`). This selection enables two controlled comparisons: (i) frontier versus efficiency-tier models within the same vendor, and (ii) consistency of feasibility judgments across vendors. All models are prompted to produce a feasibility label and a brief justification using identical task instructions.

Datasets: Our study focuses on benchmarks where feasibility judgments depend on the relationship between a hypothesis and structured scientific evidence, rather than on surface-level factual correctness or citation matching. We therefore select datasets whose evidence structure supports feasibility reasoning under different levels of uncertainty. MATTER-OF-FACT dataset (Jansen et al., 2025) closely matches our formulation because each claim is paired with a source paper whose experiments and outcomes may support, contradict, or only partially address the hypothesis. Many claims depend on experimental scope, intervention strength, or measurement conditions, reflecting realistic feasibility scenarios in which a claim may be plausible but not empirically established. REASONS dataset (Saxena et al., 2025) provides a complementary positive-feasibility setting. Each instance corresponds to a scientific statement explicitly supported by cited literature, allowing us to examine whether models correctly recognize feasibility when valid evidence exists.

Data leakage control. To reduce pretraining leakage, we apply a model-specific post-cutoff filtering rule whenever a provider-documented knowledge cutoff is available. For MATTER-OF-FACT, this filtering is based on the benchmark's claim-level temporal metadata (`published_date` / `exclude_date`). In practice, the full MATTER-OF-FACT test split is naturally post-cutoff for GPT-4o (Oct. 2023), whereas later cutoffs such as GPT-5.1 (Sep. 2024) and Gemini-2.5-Pro/Flash (Jan. 2025) require narrower claim-specific subsets from the 2024–2025 test data. This protocol reduces, but does not eliminate, leakage risk, since papers may have circulated earlier as preprints, abstracts, or secondary summaries, and provider-side training details are not fully transparent.

Controlled Evidence Construction. For instances with an associated source paper, we extract two structured evidence components: (i) experiment descriptions $\mathcal{E}^\star$ and (ii) reported outcomes $\mathcal{O}^\star$. Extraction is performed once per instance and reused across all models. We do not provide the full paper text to avoid trivial claim matching and to isolate the effects of experimental design and outcome information. The same extracted evidence is used across all settings (H, H+E, H+O, H+E+O).

Extraction audit. Because all evidence-conditioned evaluations depend on the extracted experiments and outcomes, we manually audited a targeted sample of extraction outputs against their source papers. The audit focused on two criteria: *coverage*, i.e., whether the extracted items capture

Model	R(%)	Scenario	MOF				REASONS	
			Acc	$F1_{\text{macro}}$	*MCC*	*Rouge*	*Acc*	$F1_{\text{macro}}$
gpt-5.1	0	**H**	0.68±0.03	0.67±0.03	0.42±0.05	0.21±0.01	0.84±0.01	0.46±0.00
	50	**H+E**	**0.67±0.04**	0.65±0.04	**0.41±0.06**	0.19±0.00	0.85±0.04	0.44±0.01
		H+O	**0.67±0.05**	**0.67±0.05**	0.35±0.10	**0.21±0.00**	**0.91±0.02**	**0.47±0.01**
		H+E+O	0.65±0.05	**0.67±0.05**	0.32±0.09	0.20±0.00	0.90±0.02	**0.47±0.01**
	100	**H+E**	**0.70±0.05**	**0.69±0.05**	**0.44±0.08**	0.19±0.00	0.84±0.02	0.45±0.01
		H+O	0.66±0.04	0.66±0.04	0.33±0.07	**0.21±0.01**	0.92±0.01	**0.47±0.00**
		H+E+O	0.66±0.03	0.66±0.03	0.33±0.06	0.20±0.01	**0.93±0.02**	**0.47±0.01**
gpt-4o	0	**H**	0.60±0.04	0.55±0.04	0.32±0.02	0.22±0.00	0.79±0.02	0.44±0.01
	50	**H+E**	0.52±0.04	0.44±0.04	0.25±0.04	0.18±0.01	**0.99±0.01**	**0.49±0.00**
		H+O	**0.61±0.03**	**0.64±0.03**	**0.26±0.05**	**0.22±0.01**	0.93±0.01	0.48±0.00
		H+E+O	**0.61±0.04**	0.63±0.03	0.25±0.07	**0.22±0.01**	0.97±0.01	**0.49±0.00**
	100	**H+E**	0.55±0.04	0.47±0.03	0.26±0.03	0.17±0.00	**0.98±0.01**	**0.50±0.00**
		H+O	**0.64±0.04**	**0.64±0.04**	**0.30±0.07**	**0.22±0.00**	0.93±0.01	0.48±0.00
		H+E+O	**0.64±0.03**	**0.64±0.03**	**0.30±0.06**	0.21±0.00	0.96±0.01	0.49±0.00
gem-pro	0	**H**	0.67±0.03	0.65±0.03	0.42±0.03	0.22±0.01	0.87±0.04	0.47±0.01
	50	**H+E**	0.48±0.04	0.43±0.03	0.11±0.04	0.20±0.01	0.90±0.05	**0.47±0.01**
		H+O	**0.69±0.05**	0.67±0.05	**0.40±0.09**	**0.22±0.01**	0.89±0.07	0.46±0.02
		H+E+O	0.66±0.06	**0.68±0.06**	0.34±0.12	**0.22±0.01**	**0.91±0.07**	0.46±0.02
	100	**H+E**	0.53±0.04	0.43±0.04	0.17±0.04	0.19±0.01	**0.91±0.05**	**0.47±0.01**
		H+O	**0.66±0.04**	**0.66±0.04**	**0.33±0.08**	**0.22±0.01**	**0.91±0.13**	0.45±0.04
		H+E+O	**0.66±0.04**	**0.66±0.04**	**0.33±0.07**	**0.22±0.00**	**0.91±0.11**	0.45±0.03
gem-flash	0	**H**	0.64±0.04	0.60±0.03	0.39±0.05	0.21±0.00	0.77±0.02	0.42±0.01
	50	**H+E**	0.53±0.03	0.51±0.03	0.23±0.02	0.19±0.01	**0.90±0.04**	**0.46±0.01**
		H+O	**0.67±0.04**	**0.67±0.04**	**0.35±0.07**	**0.22±0.01**	0.86±0.03	0.45±0.01
		H+E+O	0.66±0.05	**0.67±0.05**	**0.35±0.09**	0.21±0.01	0.89±0.03	**0.46±0.01**
	100	**H+E**	0.57±0.04	0.50±0.03	0.27±0.03	0.18±0.01	**0.91±0.02**	**0.47±0.00**
		H+O	**0.66±0.04**	**0.66±0.04**	**0.33±0.07**	**0.22±0.01**	0.87±0.02	0.46±0.00
		H+E+O	0.65±0.04	0.65±0.04	0.32±0.07	0.21±0.01	0.89±0.01	**0.47±0.00**
grok	0	**H**	0.65±0.04	0.63±0.05	0.36±0.06	0.22±0.01	0.88±0.02	0.47±0.01
	50	**H+E**	0.55±0.03	0.49±0.03	0.28±0.06	0.18±0.00	**0.93±0.00**	**0.48±0.00**
		H+O	**0.68±0.04**	**0.66±0.04**	**0.36±0.09**	**0.20±0.00**	0.87±0.01	0.47±0.00
		H+E+O	0.66±0.06	**0.66±0.06**	0.33±0.12	0.19±0.00	0.92±0.02	0.47±0.00
	100	**H+E**	0.57±0.04	0.50±0.04	0.23±0.06	0.18±0.00	**0.93±0.01**	**0.48±0.00**
		H+O	**0.66±0.04**	**0.66±0.04**	**0.33±0.07**	**0.20±0.01**	0.88±0.02	0.47±0.00
		H+E+O	**0.66±0.04**	**0.66±0.04**	0.32±0.08	0.19±0.01	**0.93±0.02**	0.47±0.01

Table 1: Feasibility results under controlled reveal levels. k_1 and k_2 denote the percentages of experiments and outcomes revealed, and $R(\%) \in \{100, 50, 0\}$ denotes the reveal level. For each R, scenarios correspond to: **H** : $(k_1, k_2) = (0, 0)$; **H+E**: $(k1, k2) = (R, 0)$; **H+O**: $(k1, k2) = (0, R)$; **H+E+O**: $(k1, k2) = (R, R)$. For MATTER-OF-FACT (MOF) dataset, we report $Acc/F1_{macro}/MCC$ for label prediction and ROUGE for rationale similarity (higher is better). Also note that each result averages 5 runs over the same 615 samples across all models (gpt-5.1, gpt-4o, Gemini-2.5-Pro/gem-pro, Gemini-2.5-Flash/gem-flash, Grok-4.1-fast/grok). See Figure 4 (in Appendix) for a compact visual summary of the main trends.

the key experiments and outcomes needed for feasibility judgment, and *precision*, i.e., whether the extracted content is faithful to the source without adding unsupported details. Representative audit examples are provided in Table 2 (Appendix A.2).

Controlled prompting design. Our goal is diagnostic evaluation rather than workflow engineering. We therefore use a minimal, fixed prompting setup so that performance differences can be attributed to the evidence provided rather than to added reasoning scaffolds or verification pipelines. Keeping the prompt structure constant across models, datasets, and evidence conditions helps isolate the effect of experiments and outcomes on feasibility judgments.

Partial-Information Stability Protocol: To evaluate robustness under incomplete evidence, we progressively remove portions of the extracted experiments and outcomes. For the previously defined reveal fractions k_1 and k_2 over $\{0, 0.5, 1.0\}$, we sample subsets uniformly without replacement for each configuration and evaluate them over five independent runs. Results are averaged to reduce sampling variance and to reveal non-monotonic degradation.

4 Results and Analysis

Table 1 summarizes feasibility performance under hypothesis-only and controlled evidence settings and underpins the analyses that follow.

RQ1: Feasibility from Internal Knowledge: Under the hypothesis-only setting (H), all models achieve non-trivial feasibility performance, indicating that parametric knowledge alone supports coarse feasibility judgments. However, performance variance across models is substantial, and justifications frequently rely on generic plausibility rather than explicit experimental constraints. This establishes H as a meaningful but brittle baseline: models can form priors about feasibility, but these priors are weakly grounded and sensitive to subsequent evidence.

RQ2: Two Consistent Patterns in How Experiments and Outcomes Affect Feasibility

Finding 1: Outcome evidence is more stabilizing than experiment descriptions. Providing outcomes (H+O) improves or preserves feasibility judgments more reliably than experiments alone (H+E). In contrast, H+E frequently degrades performance relative to H, particularly under partial evidence. This result is notable because experiments are often assumed to be the primary unit of scientific reasoning. Our findings suggest that, for current LLMs, experiment descriptions without results introduce ambiguity that models do not resolve reliably.

Finding 2: More evidence is not strictly better. In several settings, H+E+O does not outperform H+O, and in some cases underperforms it. This indicates that models do not consistently integrate experiments and outcomes compositionally; instead, additional context can interfere with decision-making.

RQ3: Robustness Under Partial Evidence: We next analyze robustness using progressive evidence removal. If models aggregated evidence coherently, performance would degrade smoothly as experiments or outcomes are removed.

To see this pattern clearly, we compare the hypothesis-only baseline (H) against the 50% and 100% reveal conditions for H+E, H+O, and H+E+O in Table 1 (also shown in Appendix Figure 4). The results show that performance does not drop smoothly as evidence is held back: in several cases, the 50% condition performs worse than the 100% version, and sometimes even worse than using the hypothesis alone. This directly shows that partial scientific context can mislead the model, not just give it less to work with.

Finding 3: Degradation is non-monotonic and sometimes adversarial: Across models, we observe non-monotonic degradation: intermediate reveal levels (e.g., 50%) often perform worse than both full evidence and no evidence. Moreover, a substantial fraction of partial-evidence configurations fall below the H baseline. This behavior demonstrates that partial scientific context can be *actively misleading*, not merely insufficient. Such effects are invisible in standard full-context evaluations.

Explaining the Failure Modes: Our results point to three systematic limitations in current LLM-based feasibility assessment. (a) *Surface alignment over evidential validity:* Models often overweight lexical or topical alignment between hypotheses and experiment descriptions, even when critical variables (e.g., intervention strength, endpoints, controls) do not match. Outcome statements mitigate this failure by imposing explicit directional constraints that are harder to rationalize away. (b) *Lack of evidence relevance gating:* When provided with mismatched or incomplete evidence, models rarely question whether the evidence actually tests the hypothesis. Instead, they attempt to reconcile any provided context, leading to brittle or incorrect feasibility judgments. (c) *Anchoring under partial information:* Under partial evidence, models appear to anchor on the available fragment and overcommit, rather than reverting to uncertainty. This explains why partial evidence can degrade performance more than no evidence at all.

5 Conclusion

Our findings suggest that current LLMs treat feasibility as surface-level classification rather than as evidence-conditioned judgment. Improving feasibility assessment will require mechanisms that explicitly model evidence relevance, uncertainty, and variable alignment, rather than relying on additional context alone. Importantly, our controlled evidence framework reveals failure modes that are not detectable in standard claim-verification or retrieval-augmented settings, highlighting the need for diagnostic evaluations when deploying LLMs in scientific workflows.

Limitations

Protocol granularity. Our stability analysis uses three controlled reveal levels ($k_1, k_2 \in \{0, 0.5, 1.0\}$) with uniform subsampling to isolate when partial evidence helps versus misleads. Finer-grained reveal schedules and structured omission regimes (e.g., retaining only the most decision-critical outcomes) are natural extensions.

Residual pretraining leakage risk. Although

we use post-cutoff paper filtering for models with documented knowledge cutoffs, leakage risk is not zero. Some source papers may have circulated earlier through preprints, abstracts, or other public summaries, and provider training corpora are not fully transparent. Accordingly, our evaluation should be interpreted as reducing leakage risk under a conservative post-cutoff protocol, rather than proving the absence of memorized exposure.

Scope of models and benchmarks. We study a small set of widely used API-accessible LLMs on datasets that provide claim-level feasibility labels and associated source evidence. Our conclusions are therefore best interpreted as characterizing *these* models and *these* evidence formats under a controlled-evidence evaluation, rather than as an exhaustive survey of scientific domains or model families (e.g., open-weight systems).

Task and metric choices. We model feasibility as a binary decision to support controlled comparisons across evidence conditions while holding the prediction task fixed. This design is intentional: our goal is not to fully model the uncertainty, conditionality, or partial support that often characterize real scientific judgment, but to isolate how LLM predictions change as evidence availability and type vary. This formulation is also aligned with the annotation schema of the datasets used in our study, which provides claim-level FEASIBLE/INFEASIBLE labels. Introducing additional categories, such as partially supported, uncertain, or conditionally feasible, would add additional sources of variation that are difficult to disentangle from the main factor under study, namely the evidence provided to the model. In this setting, binary labels are especially useful because label changes provide an unambiguous signal of evidence sensitivity under controlled evidence conditions. At the same time, we acknowledge that binary feasibility does not capture the full granularity of scientific reasoning and therefore limits the generalizability of our conclusions to richer expert decision settings. Extending the framework to multi-level or probabilistic feasibility judgments would be an important direction for future work, but would require annotations that are not available in the datasets used here.

Controlled prompting. We maintain consistent task instructions across all models and datasets, ensuring that any changes in predictions are attributable to the evidence provided, rather than to variations in prompts.

Ethical considerations

This work analyzes how Large Language Models (LLMs) assess scientific feasibility under controlled evidence settings (hypothesis-only vs. adding extracted experiments and/or outcomes, including partial reveal). Our study uses publicly available datasets and papers and does not involve human subjects or personally identifiable information. We emphasize that feasibility predictions and rationales are not authoritative scientific judgments: models may be brittle under incomplete or mismatched evidence and may produce confident but ungrounded explanations. We also acknowledge that LLM outputs can reflect biases from pretraining and the covered scientific domains, and we do not make normative claims about model decisions. Our goal is diagnostic—to characterize when evidence helps or harms—and to support more reliable evaluation and interpretation for domain experts, rather than to enable automated acceptance/rejection of scientific claims.

Acknowledgments

This material is based on research that is in part supported by the DARPA for the SciFy program under agreement number HR00112520301. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of DARPA or the U.S. Government.

References

Steffen Eger, Yong Cao, Jennifer D'Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, Chenghua Lin, Nafise Sadat Moosavi, Wei Zhao, and Tristan Miller. 2025. Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation. *Preprint*, arXiv:2502.05151.

Peter Jansen, Samiah Hassan, and Ruoyao Wang. 2025. Matter-of-fact: A benchmark for verifying the feasibility of literature-supported claims in materials science. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4090–4102, Suzhou, China. Association for Computational Linguistics.

Quanliang Liu, Maciej P. Polak, So Yeon Kim, MD Al Amin Shuvo, Hrishikesh Shridhar Deodhar, Jeongsoo Han, Dane Morgan, and Hyunseok Oh. 2025. Beyond designer's knowledge: Generating materials design hypotheses via a large language model. *Acta Materialia*, 297:121307.

Seyedali Mohammadi, Bhaskara Hanuma Vedula, Hemank Lamba, Edward Raff, Ponnurangam Kumaraguru, Francis Ferraro, and Manas Gaur. 2025. Do LLMs adhere to label definitions? examining their receptivity to external label definitions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32380–32393, Suzhou, China. Association for Computational Linguistics.

Yumi Nakamura, Hiroshi Watanabe, Aiko Tanaka, Masato Yasui, Jun Nishihira, and Norihito Murayama. 2020. Effect of increased daily water intake and hydration on health in japanese adults. *Nutrients*, 12(4).

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Delip Rao, Weiqiu You, Eric Wong, and Chris Callison-Burch. 2025. NSF-SciFy: Mining the NSF awards database for scientific claims. In *Proceedings of The 5th New Frontiers in Summarization Workshop*, pages 183–198, Hybrid. Association for Computational Linguistics.

Yash Saxena, Deepa Tilwani, Ali Mohammadi, Edward Raff, Amit Sheth, Srinivasan Parthasarathy, and Manas Gaur. 2025. Attribution in scientific literature: New benchmark and methods. *Preprint*, arXiv:2405.02228.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. Large language models for automated open-domain scientific hypotheses discovery. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13545–13565, Bangkok, Thailand. Association for Computational Linguistics.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From automation to autonomy: A survey on large language models in scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17733–17750, Suzhou, China. Association for Computational Linguistics.

A Appendix

A.1 Implementation Details

All models are accessed via their public APIs with a low decoding temperature (0.1) to reduce stochasticity. For each setting, we use a single prompt template per model family and apply it unchanged across datasets. We include abstracted templates for (i) extracting experiments and outcomes from the source paper (Figure 2) and (ii) feasibility prediction in the H+E+O setting (Figure 3), with outputs constrained to structured JSON. For reproducibility, we use five fixed random seeds (42, 123, 456, 789, 101112) for sampling and report results aggregated across seeds.

Abstracted Prompt for Extracting Tests (Experiments + Outcomes)

Input: `{hypothesis}`, `{paper_text}`.

Task: Extract the most relevant **experiments** and **outcomes** needed to judge whether the hypothesis is feasible.

Rules:

- **Experiments** ($\leq$10): what was tested and what was measured (1–2 sentences each).
- **Outcomes** ($\leq$5): key results (direction + important limitation) (1–2 sentences each).
- Use **only** evidence reported in the provided paper text.

Output: Return **ONLY** valid JSON:

```
{
  "experiments": ["..."],
  "outcomes": ["..."]
}
```

Figure 2: Abstracted prompt for extracting experiments and outcomes from the source paper.

Abstracted Prompt for Feasibility (Hypothesis + Experiments + Outcomes)

Input: `{hypothesis}`, `{experiments}`, `{outcomes}`.

Task: Decide whether the hypothesis is FEASIBLE or INFEASIBLE given the provided experiments and outcomes.

Rules:

- Base the decision on the provided evidence (do not add new experiments or results).
- If evidence is insufficient or conflicting, choose the best-supported label and state uncertainty in the explanation.
- **Explanation** (1–3 sentences): cite specific experiments/outcomes.

Output: Return **ONLY** valid JSON:

```
{
  "decision": "feasible" or "infeasible",
  "explanation": "1--3 sentences referencing
                  the experiments/outcomes"
}
```

Figure 3: Abstracted prompt for feasibility prediction given a hypothesis, experiments, and outcomes.

A.2 Manual Audit of Extracted Experiments and Outcomes and Stability Curves

Table 2 shows representative examples from our manual audit of extracted experiments and outcomes against their source papers. For each case, we report an abstracted summary of the extracted experiments and outcomes, together with whether the extraction captured the key evidence needed for feasibility judgment (*coverage*), whether it remained faithful to the source without unsupported additions (*precision*), and the main audit note or failure mode.

Case	Dataset	Claim / Hypothesis	Abstracted extracted experiments	Abstracted extracted outcomes	Coverage	Precision	Main audit note / failure mode
1	MoF	The charge transfer resistance in lithium-ion batteries follows a U-shaped pattern with respect to state-of-charge, with higher resistances at very low and very high SOC values and lower resistances in the middle SOC range. (Claim ID: 2412.10896v3_4_T)	• Full-cell cycling and impedance measurements across SOC. • Symmetric-cell analyses to isolate electrode-specific effects. • Temperature- and rate-dependent tests to separate charge-transfer contributions.	• Charge-transfer resistance shows a clear mid-SOC minimum. • Resistance rises at both low and high SOC. • The trend persists across temperature and aging, with graphite dominating at low SOC and NMC at high SOC.	Pass	Pass	The extraction captures the key electrochemical measurements and the central U-shaped finding, while preserving the mechanistic interpretation rather than reducing it to a generic resistance trend.
2	MoF	Higher growth temperatures for Ta films on sapphire lead to improved structural and DC electrical properties but paradoxically result in worse microwave performance due to interface-related loss mechanisms. (Claim ID: 2412.16730v1_10_T)	• Ta films were grown on sapphire at multiple substrate temperatures. • Structural characterization used XRD, TEM, and AFM. • Electrical and superconducting properties were measured with transport and resonator experiments.	• Higher growth temperature improves crystallinity, surface quality, and DC/superconducting transport. • Microwave resonator internal quality factor decreases at higher growth temperature. • The degradation is attributed to lossy Ta/sapphire interfacial states or layers.	Pass	Pass	The extraction covers both the improvement in structural/DC properties and the degradation in microwave behavior, which is essential to the claim's paradoxical conclusion.
3	MoF	Substituting Si with Ge in $LaRu_3Si_2$ increases the superconducting transition temperature from 6.6 K to 7.1 K at $x = 0.07$ Ge content. (Claim ID: 2503.22477v2_0_T)	• A $LaRu_3Si_{2-x}Ge_x$ substitution series was synthesized. • Samples were characterized with XRD, resistivity, magnetization, and specific heat. • Normal-state transport was examined across composition.	• Ge substitution is structurally incorporated successfully. • Superconducting T_c does not increase with Ge substitution. • Bulk superconducting signatures shift to lower temperature, consistent with disorder-induced suppression.	Pass	Pass	This is a strong negative-feasibility case: the extraction directly addresses the stated composition and temperature claim and preserves the contradiction clearly.
4	REA	Hardware-efficient ansatzes that ignore chemical information are unlikely to support large molecules effectively. (Claim ID: 2105.07127v1_25)	• The paper compares hardware-efficient, UCCSD, QCC, and other chemically informed ansatzes. • Experiments evaluate molecular VQE accuracy, circuit depth, and optimization behavior. • Scaling with molecule size is also examined.	• Chemically informed ansatzes achieve better energy accuracy at lower depth. • These ansatzes scale more favorably to larger molecules. • Hardware-efficient ansatzes often require impractical depth and optimization effort as molecule size grows. (Paper ID: arXiv:1803.11173)	Pass	Partial	The extraction supports the core claim well, but it is somewhat broader than the exact statement span and includes additional conclusions about optimization and scaling beyond the most local claim wording.
5	REA	Gumbel-Top-k uses a softmax-based continuous relaxation of argmax/top-k and achieves differentiability via reparameterization. (Claim ID: 2312.14474v1_22)	• The paper studies Gumbel-Top-k as a differentiable subset-selection / top-k operator. • It compares the method with non-differentiable and REINFORCE-style baselines. • Experiments examine training behavior and task performance on ranking/subset-selection tasks.	• The method enables end-to-end gradient-based training through a reparameterized softmax relaxation. • It reduces gradient variance relative to score-function estimators. • It performs competitively but introduces a temperature-dependent approximation bias.	Pass	Pass	The extraction preserves both the main methodological benefit and the caveat about approximation bias, making it a clean example of faithful evidence abstraction.

Table 2: Manual audit summary of extracted experiments and outcomes. Coverage indicates whether the extracted evidence includes the main experiments and outcomes needed for feasibility judgment. Precision indicates whether the extraction is faithful to the source paper without unsupported additions. Claim ID denotes the identifier of the benchmark claim/example, which is associated with a source paper. MoF: MATTER-OF-FACT, REA: REASONS.

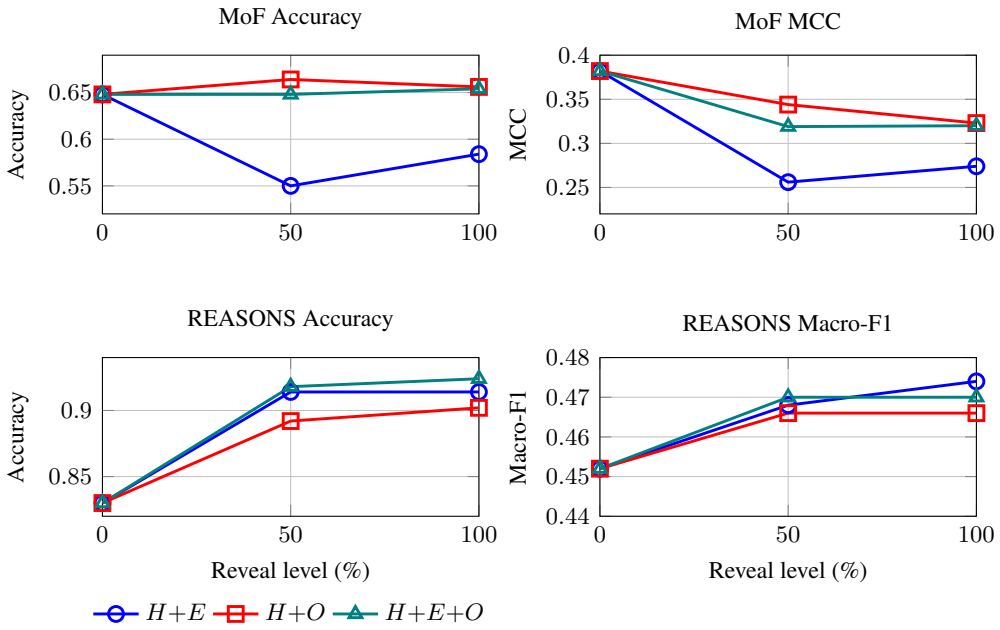

Figure 4: Reveal-level stability curves derived from Table 1, where each point is averaged over the five evaluated models for the corresponding metric and reveal setting. Each curve starts from the shared hypothesis-only baseline H at 0% reveal and shows performance under 50% and 100% reveal for $H{+}E$, $H{+}O$, and $H{+}E{+}O$. These curves show that robustness depends strongly on evidence type: on MoF, $H{+}E$ is the most brittle condition, especially under partial reveal, whereas outcome-based settings are more stable; on REASONS, evidence augmentation is generally beneficial relative to H, though the gains are not uniform across settings.