MATTER-OF-FACT: A Benchmark for Verifying the Feasibility of Literature-Supported Claims in Materials Science

Peter Jansen[1,2], **Samiah Hassan**[1], **Ruoyao Wang**[1]
[1]University of Arizona, [2]Allen Institute for Artificial Intelligence
pajansen@arizona.edu

Abstract

Contemporary approaches to assisted scientific discovery use language models to automatically generate large numbers of potential hypothesis to test, while also automatically generating code-based experiments to test those hypotheses. While hypotheses can be comparatively inexpensive to generate, automated experiments can be costly, particularly when run at scale (i.e. thousands of experiments). Developing the capacity to filter hypotheses based on their feasibility would allow discovery systems to run at scale, while increasing their likelihood of making significant discoveries. In this work we introduce MATTER-OF-FACT, a challenge dataset for determining the feasibility of hypotheses framed as claims, while operationalizing feasibility assessment as a temporally-filtered claim verification task using backtesting. MATTER-OF-FACT includes 8.4K claims extracted from scientific articles spanning four high-impact contemporary materials science topics, including superconductors, semiconductors, batteries, and aerospace materials, while including qualitative and quantitative claims from theoretical, experimental, and code/simulation results. We show that strong baselines that include retrieval augmented generation over scientific literature and code generation fail to exceed 72% performance on this task (chance performance is 50%), while domain-expert verification suggests nearly all are solvable – highlighting both the difficulty of this task for current models, and the potential to accelerate scientific discovery by making near-term progress.[1]

1 Introduction

Contemporary language models are being broadly integrated into the scientific discovery pipeline. Existing systems can generate hypothesis (Si et al., 2024; Radensky et al., 2024), run experiments (Lu et al., 2024; Li et al., 2024; Jansen et al., 2025), analyze data (Majumder et al., 2025), and write or review papers (Liu and Shah, 2023; Zhou et al., 2024). A central benefit – and challenge – of these systems is that they can function at scales greater than any human scientist. For example, hypothesis generation systems might easily produce thousands of potential hypotheses (Lu et al., 2024; Jansen et al., 2025), and running experiments to test each of these would be costly and impractical – particularly in that few experiments are likely to yield positive results. In this work we investigate the task of *feasibility assessment* (e.g. O'Neill et al., 2025), or assessing whether we can filter hypothesis (expressed as claims) to those that are most likely to be feasible, and have their hypotheses confirmed. Performing well at this task would allow us to incorporate feasibility filtering in hypothesis generation systems, and potentially make more discoveries with a given (fixed) experimental budget.

Feasibility assessment is in principle quite challenging as it involves (at times) a high degree of uncertainty in predicting future results, and yet it is a task that scientists perform frequently during experiment planning stages – selecting the hypotheses that we believe are most likely to return positive results based on a combination of literature, pilot experiments or analyses (which may include empirical work, or code/simulations), and past experience. In this work we aim to investigate how well current models can perform this feasibility assessment task, and provide a benchmark to assist in improving model performance over time.

Generating data to test feasibility assessment is challenging, as (by nature) the experimental results of proposed hypotheses are as yet unknown, which makes gold labels for determining whether a hypothesis is feasible or infeasible effectively unavailable. To address this challenge, *we operationalize feasibility assessment as a temporally-filtered claim verification task* using the concept of

[1]Benchmark, models, and claim extraction system: https://github.com/cognitiveailab/matter-of-fact

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 4091–4103
November 4-9, 2025

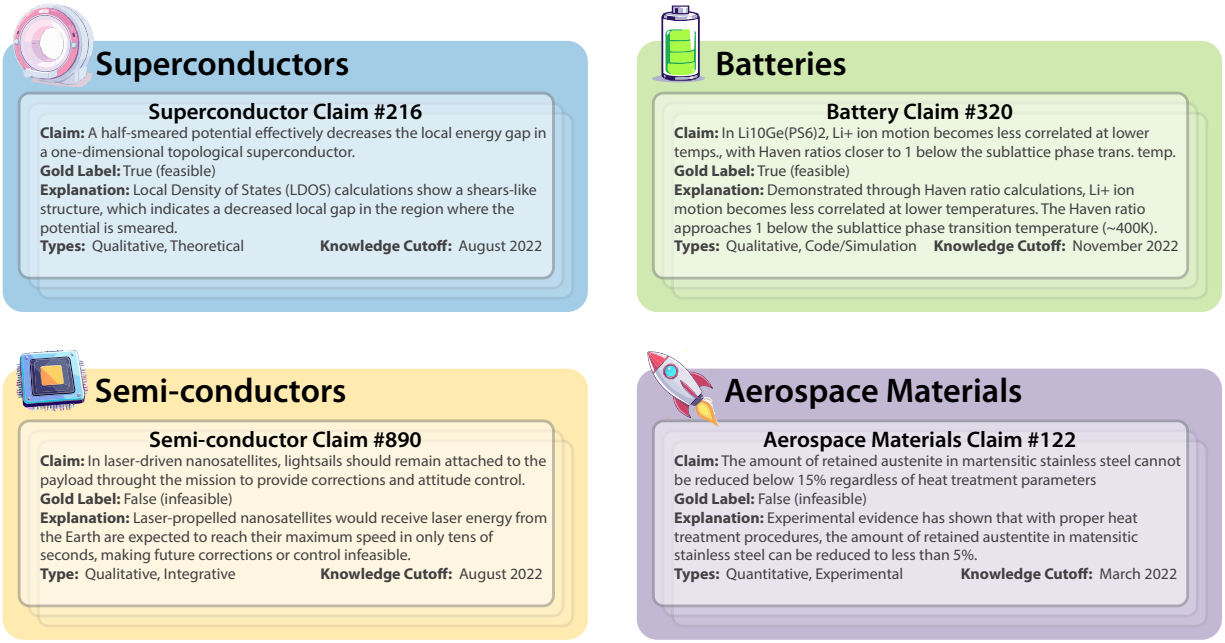

Figure 1: Examples of the four main materials science topics included in MATTER-OF-FACT, including *superconductors*, *semi-conductors*, *batteries*, and *aerospace materials*. Each claim includes the claim text, gold label *(true/feasible* or *false/infeasible*), a brief explanation and supporting facts based from the original paper the claim was sourced from, additional meta-data (such as whether the claim is *quantitative* or *qualitative*), and the knowledge cutoff date for the feasibility assessment task.

backtesting. As with conventional claim verification tasks (e.g. Thorne et al., 2018), we generate a corpus of claims extracted from recent scientific literature – however, in addition, each claim is paired with a "knowledge cut-off date", which is the date that the paper the claim was generated from was first authored. When performing the feasibility assessment task, models are allowed to use any information available before the source paper was authored to assess feasibility, essentially rewinding into the past to predict future results. In this way, models are provided with knowledge up to (for example) 2023, and must use that knowledge (through a combination of literature search, small-scale code-based experimentation, world modeling, or other methods) to predict whether genuine results (and artificially-generated infeasible results) from 2024 onward are feasible or infeasible.

The contributions of this work are:

1. We introduce MATTER-OF-FACT, a benchmark of 8.4K claims extracted from recent materials science articles in four high-impact subdomains. Each claim includes categorical information (qualitative vs quantitative, and experiment, code, or theory focused), and is paired with a knowledge cut-off date to use for the feasibility assessment task.

2. We empirically demonstrate that strong baseline models using a variety of solution methods (including retrieval-augmented generation with SEMANTICSCHOLAR, as well as evidence gathered from code-generation) across base models (GPT-4O-MINI, O4-MINI, and CLAUDE SONNET 3.7) achieve a maximum of 72% accuracy, highlighting the challenging nature of this feasibility assessment task.

3. We assess the quality of the claims both by domain expert evaluation, and by evaluating base models in a conventional claim verification task. Humans and models reach 93%+, suggesting the benchmark is of high quality.

2 Related Work

Scientific Claim Verification Datasets: The scientific claim verification task requires a model to determine whether a claim (typically extracted from a scientific paper) is true or false, either by leveraging its pretrained scientific knowledge or retrieving evidence from a corpus, with a selection of scientific claim verification benchmarks shown in Table 1. SCIFACT (Wadden et al., 2020) contains 1.4K biomedical-domain claims generated by showing citances (sentences that cite a paper and describe its contribution) to human annotators, who were

Benchmark	Domain	Claims	Source	Generation Method
SciFact (Wadden et al., 2020)	Biomed	1.4K	Paper Abstracts	Citances provided to annotators
COVID-Fact (Saakyan et al., 2021)	Biomed	4.0K	Reddit	Extract positive, generate counterclaim
SciFact-Open (Wadden et al., 2022)	Biomed	1.4K	Paper Abstracts	See SciFact
ClaimCheck (Ou et al., 2025)	ML	154	Paper Reviews	Emphasizes claim weaknesses
SciTab (Lu et al., 2023)	Comp. Sci	1.2K	Paper Tables	Compositional reasoning on tables
Matter-of-Fact (This work)	**Mat. Sci**	**8.4K**	**Paper full-text**	**Extract positive, generate infeasible**

Table 1: A comparison of claim verification datasets with Matter-of-Fact, including their domain, size, source of the information used to generate or extract claims from, and the claim generation method.

then asked to generate associated claims. Where SciFact pairs claims with a set of 5K abstracts that can be used for gathering evidence, SciFact-Open (Wadden et al., 2022) expands this evidence retrieval corpus to 500K abstracts, presenting a more challenging retrieval problem. Also in the biomedical domain, COVID-Fact (Saakyan et al., 2021) consists of over 4K claims extracted from Reddit. Lu et al. (2023) introduce SciTab, which requires verifying computer science claims centrally using tables extracted from papers. ClaimCheck (Ou et al., 2025) uses reviews of rejected NeurIPS submissions from OpenReview to build a corpus of 154 claims that emphasize identifying the weaknesses in scholarly claims. In contrast, Matter-of-Fact builds a corpus of 8.4K materials science claims for feasibility assessment that are generated from the nuanced results found in the full text of source articles (rather than abstracts), and where negative claims focus on being scientifically infeasible rather than factually incorrect.

Claim Verification Models: Our framing of feasibility detection is as temporally-filtered claim verification with a knowledge cutoff. More broadly, recent approaches to claim verification typically involve two key steps: evidence retrieval and fact checking. For the retrieval step, augmenting LLMs with retrieved documents (Izacard et al., 2022) or knowledge bases (Baek et al., 2023; Hang et al., 2024) can be effective for improving fact verification performance of models. Re^{2}G (Glass et al., 2022) extends the retrieval step with a trained reranker to achieve better retrieval performance for fact checking. Rani et al. (2023) propose a form of query expansion that generates claim-related questions as queries to retrieve supporting documents. For fact checking, some methods make use of structured knowledge representations such as knowledge graphs (Dammu et al., 2024) and first-order-logic (Wang and Shu, 2023) to organize evidence and verify facts. End-to-end systems combine the entire retrieval and verification pipeline, such as ARSJoint (Zhang et al., 2021) and SciClaims (Ortega and Gómez-Pérez, 2025). In this work we demonstrate similar retrieval-backed systems (with temporal filtering) for feasibility assessment, while also providing formal approaches based on code generation.

Scientific Discovery and Feasibility Assessment: Automated scientific discovery is frequently divided into two subfields: problem-specific methods (like AlphaFold (Jumper et al., 2021) for protein structure prediction), and problem-general methods that work across a variety of problem types. Examples of problem-specific systems in the materials science domain include GNoME (Merchant et al., 2023), a graph neural network (GNN) based method that discovered over 2.2 million new stable crystal structures, and Schmidt et al. (2023)'s method for using crystal-graph neural networks together with high-quality data for accurate stability prediction. The latter work screened 1 billion materials, discovering 150k+ stable compounds, and identified extreme-property materials like superconductors. Similarly, Chen et al. (2024) combine machine learning models with traditional physics-based models to discover compounds to which can potentially serve as solid electrolytes. These problem-specific methods can be applied to feasibility assessment by predicting highly specific properties of unknown materials. Matter-of-Fact works to bridge the gap between problem-specific methods and problem-general methods by providing a large set of claims across 4 broad and high-impact areas of materials science, each of which is likely to benefit from a variety of problem-specific methods to arrive at accurate feasibility assessments. As we empirically demonstrate in our modeling results, because Matter-of-Fact nominally requires a large set of capacities to solve, it is challenging benchmark for measuring a general capacity to assess feasibility over broad subdomains.

3 Dataset

The MATTER-OF-FACT benchmark consists of 8.4K claims extracted from the full-text of materials science articles. The extraction and validation process is described below, with example claims shown in Figure 1.

Inclusion Criteria: We assembled a corpus of recent publicly-available materials science domain articles by crawling Arxiv for all papers within the MATERIALS SCIENCE and SUPERCONDUCTIVITY topics submitted on or after January 2022, resulting in a total of 24K articles. Articles were then filtered based on specific inclusion criteria. First, articles that were not licensed using a specific permissive license (CREATIVE COMMONS-BY ATTRIBUTION-4.0) were removed. Second, to prevent having to use a PDF-TO-TEXT conversion pipeline (which can have limited quality on complex tables, chemical formulas, mathematics, and other artifacts found within materials science articles), we further filtered to include only articles with LaTeX source available. Papers with long source (>30k tokens) were also removed (approximately 16% of articles). After initial filtering, 4.2K articles remained. Our focus in this work is specifically in four high-impact subdomains: superconductors, semi-conductors, batteries, and aerospace materials. To identify articles within these topics, we performed topic labeling of each abstract using GPT-4O-MINI with a prompt that emphasized identifying articles within these 4 focus areas. We then sampled 500 total articles (125 from each topic) to use for claim generation.

Initial Claim Generation: Claims were generated by providing the full-text (LaTeX source) of each paper in a prompt, together with task instructions and JSON output format requirements. The model was instructed to generate matched pairs of claims – one true, and one that was clearly false or infeasible – and for each claim, to provide a list of supporting evidence, followed by an overall explanation as to why the evidence supports or refutes the claim.[2]

Claims were instructed to be stand-alone, and not make reference to the paper in the claim text (i.e. "*Table 4* claims the boiling point of Material X is..."), so that they could be (in principle) solved without reference to the original source paper. Negative claims were instructed to be false or clearly infeasible (but not overly so), and not simply claims for which no evidence was available. Similarly, negative claims were instructed to use balanced language so as not to give away their true or false nature by particulars of wording, such as through use of negation markers (i.e. "Material X does *not* have..."), and to instead use neutral framings. In addition to the above constraints, claims were explicitly asked to be authored on two dimensions. The first asks for claims to specifically test either *qualitative* knowledge (e.g. "*In Situation X, Phenonemon Y helps Material Z maintain its superconductivity*"), or *quantitative* knowledge (e.g. "*Material X superconducts at 77 Kelvin*"). Second, claims were asked to be authored cross 4 main types: those that focus on *experimental results*, *code/simulation results*, *theoretical results*, or *integrative methods* across types.

Balanced Temporal Sets: Claims were temporally sorted into those from papers first appearing on Arxiv in 2022 (for training), those in 2023 (for validation), and the most recent claims from papers submitted between 2024 and April 2025 (for testing). For each set, we filtered claims such that equal numbers of true and false claims were present, to achieve a baseline (random chance) performance of 50%. Claims were also balanced such that equal numbers within the *experimental, code, theory,* and *integrative* categories appear within a given set. The final dataset includes a total of 8.4K claims, distributed as 1.4K claims for training, 2.5K for validation, and 4.4K for testing.

Domain Expert Validation: To measure the quality of the claim generation process, a domain expert with a graduate degree in materials science was given each claim and its source paper, and independently asked to determine the validity of the claim. This was a challenging task, because the claims span broad areas of materials science that would be unusual for any single individual to have expertise within. The domain expert performed this task for 100 claims from the test set, and initially agreed in 93% of cases (while noting a further 3% of claims appeared to not meet criteria, such as explicitly referencing the original source paper). They were then provided with the LLM-generated labels and explanations, and asked to resolve disagreements (either LLM errors, or human error), noting that nearly all errors were a result of the

[2] Scientific articles tend to express positive claims rather than negative claims. We follow the approach of Saakyan et al. (2021) to first extract positive claims, then automatically generate negative claims from these positive references.

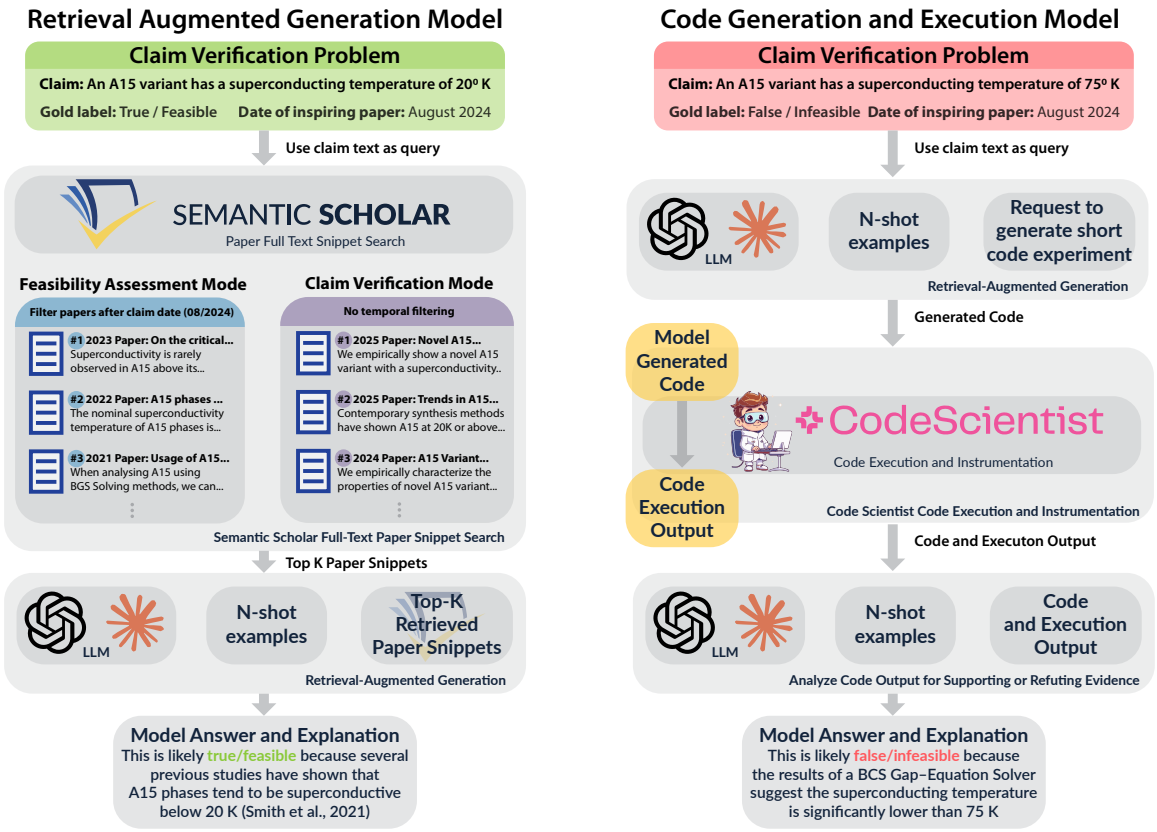

Figure 2: Flow diagrams for two models: the retrieval-augmented generation (RAG) model that retrieves snippets from the full-text of papers using SEMANTICSCHOLAR (left), and a code generation model that executes PYTHON code and examines the output using CODESCIENTIST.

domain expert missing difficult-to-find evidence on their first attempt, and ultimately reaching 99% agreement after this resolution process. This empirically suggests that the overall data quality is high (96% after discounting data not meeting generation criteria). Before resolution, interrater agreement using Cohen's Kappa (Cohen, 1960) was $\kappa = 0.86$, or strong agreement (McHugh, 2012), suggesting the claims are highly objective.

4 Baseline Models

We evaluate performance on the MATTER-OF-FACT dataset using a selection of baseline models described below. Models are provided with the text of the claim, and must predict a binary label (true/feasible, or false/infeasible), as well as provide a brief explanation for their reasoning. All models investigated in this work use in-context learning (ICL), and are characterized across three common base models at different price/performance points, including GPT-4O-MINI, O4-MINI, and CLAUDE SONNET 3.7. Our retrieval-augmented generation and code-generation models are shown in Figure 2.

4.1 Feasibility Assessment Models

Chain-of-Thought (CoT), ICL, Reflection: The language model is provided with a prompt that includes the claim, and a request to think and/or plan before responding in the style of Chain-of-Thought (Wei et al., 2022). We also include two variations of this model. The first includes a *20-shot* in-context learning example (Brown et al., 2020), using 20 claim problems (together with their supporting facts and explanations) drawn from the training set, including balanced numbers of true and false claims. The second includes adding a reflection step (Madaan et al., 2023) where, after the initial generation, the model then reflects on its response, then provides a final answer and explanation for the reasoning behind that answer.

Retrieval Augmented Generation (RAG): Using the claim text as a query, the model first retrieves the *top K* matching full-text snippets from scholarly scientific articles using the SEMANTICSCHOLAR API (Kinney et al., 2023), where each snippet generally takes the form of a span of text (approximately 500 words in length) from an article indexed by SEMANTICSCHOLAR that most closely

Model	Overall Accuracy	Accuracy by Category								Cost (per 1k)
		True	False	Qual.	Qnt.	Exp.	Code	Ther.	Int.	
RANDOM BASELINE	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0
GPT-4O-MINI										
CHAIN-OF-THOUGHT (COT)	0.65	0.89	0.40	0.70	0.57	0.61	0.62	0.66	0.69	$1
COT + 20-SHOT ICL	0.66	0.58	0.75	0.71	0.60	0.64	0.64	0.65	0.73	$2
COT + 20-SHOT ICL + REFLECTION	0.67	0.59	0.74	0.71	0.60	0.64	0.63	0.66	0.73	$3
COT + ICL + RAG (SEMANTICSCHOLAR)	0.68	0.82	0.55	0.74	0.61	0.65	0.66	0.69	0.74	$3
COT + ICL + CODE (CODESCIENTIST)	0.64	0.61	0.66	0.70	0.56	0.60	0.60	0.65	0.71	$4
O4-MINI										
CHAIN-OF-THOUGHT (COT)	0.58	0.62	0.54	0.66	0.47	0.52	0.56	0.58	0.67	$4
COT + 20-SHOT ICL	0.71	0.71	0.70	0.76	0.63	0.68	0.67	0.72	0.76	$15
COT + 20-SHOT ICL + REFLECTION	0.71	0.72	0.69	0.76	0.63	0.68	0.68	0.71	0.76	$30
COT + ICL + RAG (SEMANTICSCHOLAR)	0.71	0.60	0.82	0.75	0.65	0.68	0.67	0.72	0.76	$27
COT + ICL + CODE (CODESCIENTIST)	0.68	0.66	0.71	0.73	0.62	0.67	0.65	0.69	0.73	$34
CLAUDE-SONNET 3.7										
CHAIN-OF-THOUGHT (COT)	0.70	0.79	0.61	0.76	0.62	0.66	0.67	0.71	0.77	$9
COT + 20-SHOT ICL	0.72	0.87	0.56	0.78	0.63	0.68	0.66	0.73	0.79	$44
COT + 20-SHOT ICL + REFLECTION	0.66	0.87	0.45	0.71	0.58	0.64	0.60	0.67	0.72	$87
COT + ICL + RAG (SEMANTICSCHOLAR)	0.71	0.64	0.77	0.76	0.63	0.69	0.63	0.74	0.77	$76
COT + ICL + CODE (CODESCIENTIST)	0.63	0.75	0.50	0.66	0.58	0.61	0.60	0.63	0.66	$173

Table 2: Model performance on the **feasibility assessment** task, including overall performance, as well as performance broken down by specific categories of feasibility assessment claim problems. *True* and *False* represent performance on problems with those gold labels. *Qual.* and *Quant.* represent performance on qualitative and quantitative problems. *Exp.*, *Code*, *Ther.*, and *Int.* represent performance on claims focusing on experimental, code/simulation, theoretical, or integrative results, respectively. Cost represents the estimated model cost per 1000 claims, in US dollars.

Model	Overall Accuracy
GPT-4O-MINI	
RAG (SEMANTICSCHOLAR (NO DATE))	0.76
ORACLE SOURCE PAPER	0.71
O4-MINI	
RAG (SEMANTICSCHOLAR (NO DATE))	0.90
ORACLE SOURCE PAPER	0.96
CLAUDE-SONNET 3.7	
RAG (SEMANTICSCHOLAR (NO DATE))	0.87
ORACLE SOURCE PAPER	1.00
HUMAN DOMAIN EXPERT	
INITIAL ASSESSMENT	0.93
AFTER RESOLVING DISAGREEMENTS	0.99

Table 3: Model performance on the **claim verification** task, using oracle models. †Note that due to the high model cost, the ORACLE SOURCE PAPER (SONNET) model is assessed on a subset of the test set.

matches the query. To prevent temporal contamination with oracle knowledge, full-text snippets are filtered such that papers authored after the source paper for a given claim are not included in the search. For example, if a claim was derived from a paper first published on Arxiv in March 2024, then only papers authored in February 2024 or before will be included in the snippet search. The top 20 matching full-text snippets (sorted by the provided relevance score) are included in the language model prompt, in a retrieval-augmented-generation paradigm (Lewis et al., 2020). The prompt for this model also includes a 20-shot ICL example, and request for chain-of-thought reasoning.

Code Generation (CODESCIENTIST): This model is performed in two stages. During the first stage, the model is provided with the claim text, and prompted to generate a code-based experiment or simulation in PYTHON that would produce useful evidence in supporting or refuting the claim. The code is then executed, and the code and execution results are provided to a second prompt with a request to generate an answer for the feasibility task as well as a supporting explanation. For code execution, we use the experiment execution portion of CODESCIENTIST (Jansen et al., 2025), which allows executing arbitrary PYTHON code in a virtual sandbox on MODAL.COM, and supports installing external supporting libraries through PIP. While this execution pipeline stores and saves out-

put streams (e.g. STDOUT/STDERR), the model explicitly prompted to save a log of its work, as well as a final list of results, which are then provided back to the model to help make its final decision. For tractability, we run CODESCIENTIST in a *highly limited form* due to its high overall cost (initially reported as $4 per experiment), which would be intractable for the size of our dataset ($\approx \$16k$ for 4K test claims). Instead of 25 debug iterations, we run CODESCIENTIST for a single iteration (without reflection), and reduce the experiment time limit from 6 hours to 10 minutes (or 31 total CPU-days across all test claims). The model is made aware of these limitations in the code generation prompt, and encouraged to design appropriately-scoped experiments and output to support the decision process.

4.2 Claim Verification Models

As a method of characterizing model performance when oracle information is available, Table 3 also provides two models that perform a *claim verification* task rather than the *feasibility assessment* task – that is, they do not have the same temporal restrictions, and are able to use data available after the source claim was authored.

RAG (Temporally Unrestricted): The retrieval-augmented generation model described above, but without temporal restrictions. For a given claim, snippets from any scientific article may be retrieved, including (potentially) the source article of the claim, or those that cite the source article.

Oracle Baseline: The language model is provided both with the claim, as well as the original source paper the claim was derived from (in the form of the paper's original LaTeX source retrieved from Arxiv) in a retrieval-augmented-generation paradigm. This baseline measures how well a model can verify the claim when provided with a source scientific article that directly speaks to that claim's validity/feasibility.

Oracle (Human Domain Expert): The domain expert evaluation, as described in Section 3.

4.3 Results

Feasibility Assessment Results: The performance of all models when evaluated in the *feasibility assessment* mode is shown in Table 2. Performance across all models ranges from 0.58 to 0.72, with the models that use the smallest (and least expensive) base model (GPT-4O-MINI) generally performing about 5 percent lower than the two more performant (and more costly) base models, O4-MINI and CLAUDE SONNET 3.7. Across models, adding features (such as in-context learning, reflection, RAG over SEMANTICSCHOLAR, or Code Generation) generally provides modest performance improvements, or does not improve performance over the CHAIN-OF-THOUGHT baseline, highlighting the difficulty of this task when using conventional solution methods, and its suitability as a challenge task. When examining performance broken down by category, we observe that while the overall performance of a given base model is similar with different features, some models are more performant at identifying true/feasible claims than they are at identifying false/infeasible claims, and vice versa. All models perform better at assessing the feasibility of *qualitative* claims than *quantitative* claims, with this difference between 11% and 19% across all models, potentially a result of quantitative claims requiring the ability (through code or other means) of verifying the feasibility of specific numerical values present in the claims. In line with this reasoning, claims that are based on *experiments* or *code/simulations* consistently achieve the lowest performance, while those based on *theoretical results* are next-highest, with *integrative* claims achieving the highest performance.

Claim Verification Results: The performance of models when evaluated in the *claim verification* mode is shown in Table 3. In this mode the models have no temporal restrictions, and may use knowledge from the source paper, or papers authored after the source paper (including those that may cite the source paper) as evidence to perform the claim verification task. These experiments serve two purposes. First, they identify an effective ceiling of how well a given base model can perform even when provided with the original source article used to create a claim, with O4-MINI and CLAUDE SONNET 3.7 capable of achieving nearly a 100% ceiling performance, while GPT-4O-MINI has more modest performance ceiling between 0.71 and 0.76. Second, these models serve as a consistency evaluation for the claim generation protocol, emphasizing that when strong models are asked to verify the labels of these automatically generated claims, they nearly always agree with the gold label. Further emphasizing this is the domain expert performance, who (when provided with the original source article), agreed with the LLM-generated label for 99%

Base Model	Knowledge Cutoff Date	Accuracy (before cutoff)	Accuracy (after cutoff)	Accuracy Δ	# Samples (before/after)
GPT-4O-MINI	Sept 2023	0.661	0.664	0.003	3236 / 5124
O4-MINI	May 2024	0.694	0.705	-0.011	5168 / 3192
CLAUDE-SONNET-3-7	Oct 2024	0.672	0.661	0.011	6130 / 2230

Table 4: Knowledge contamination analysis of base models. In this analysis, performance of the CHAIN-OF-THOUGHT + 20-SHOT ICL + REFLECTION model is shown for claims from papers that were authored before or after a given model's advertised knowledge cutoff date. Given the temporal nature of the dataset, all 8.4K claims across train, development, and test sets were included. All models show almost identical performance ($\pm 1\%$) when tested on claims from papers before or after their knowledge cutoff date, suggesting that knowledge contamination does not play a significant role in performance.

of claims after resolving disagreements.

Taken together, these results empirically demonstrate the generation quality of the feasibility claims, while also emphasizing that common models and architectures still achieve overall modest performance on the feasibility assessment task.

5 Discussion

5.1 Controlling Potential Confounds

Base-Model Contamination: A central part of the framing of our *feasibility assessment* task as a temporally-filtered *claim verification* task is that it requires models to have a minimum of contamination with knowledge beyond a given claim's knowledge cutoff date. While it is possible that techniques such as model editing and machine unlearning (Bourtoule et al., 2021; Tarun et al., 2023; Liu et al., 2025) may eventually allow the knowledge in a base model to be temporally filtered to minimize this contamination, this may have limited success in current forms (Lynch et al., 2024; Deeb and Roger, 2024; Du et al., 2024). Instead, here we aim to measure how much of the current model performance is likely due to model contamination (from, for example, the base model being trained on the source articles used to generate the claims). To measure this, we examine each base model's performance for claims extracted from papers before and after the base model's advertised training data knowledge cut-off dates. The results, shown in Table 4, show that the performance of the base models on claims from papers authored after their knowledge cutoff is *nearly identical* to the performance on claims authored by papers that are before the knowledge cutoff date. This empirically suggests that the performance of current base models on the feasibility assessment task is not due to model contamination, but due to other properties, such as their capacity for reasoning.

Collecting Claims from Scientific Papers: In this work, the claims we use for the feasibility assessment task (and the associated hypotheses underlying those claims) are collected from scientific articles. It is common that the narrative presented in an article differs from the actual scientific process that was undertaken, which typically includes failed experiments, promising initial directions that turned out to be dead ends, and other complexities that are often left out when writing a paper. As such, the claims (and associated hypotheses) extracted from papers may have a different (and more polished) character than those one naturally explores during the scientific discovery process. While it is currently difficult to quantify this potential difference, we wish to acknowledge that it may exist, and that this may ultimately affect transfer performance in models trained or evaluated on literature-derived claims (like those in MATTER-OF-FACT) to real-world discovery scenarios.

5.2 Performance Characterization

Pragmatic Ceiling Performance: While we empirically show that the feasibility of many claims can be assessed using inexpensive means, the models we demonstrate are far from achieving perfect performance on this task. Pragmatically, a model that achieves near 100% performance would be able to (with near perfect accuracy) determine whether claims are likely to be feasible or infeasible through some combination of literature search, inexpensive code-based experimentation, world modeling, and other means. Achieving 100% performance is likely impractical, as many scientific claims can only be verified with empirical work, and not with literature search or simulation, particularly for those (most impactful) scientific results that are surprising because they run counter to expectations. That being said, even though effective ceiling performance on feasibility assessment tasks is likely to be less than 100%, increasing model performance on this task even a modest amount can have practical utility for improving the efficiency of

discovery systems. As we show in APPENDIX A, for a hypothetical hypothesis generation system where 1% of the hypotheses are true, the performance of our current-best model could potentially allow discovering 60% of the true hypotheses while reducing experiment costs by 80% – a large overall budget reduction, at the cost of reducing the recall of finding true hypotheses by approximately 40%.

Limitations in Baseline Models (Retrieval): It is entirely plausible that assessing the feasibility of complex state-of-the-art scientific claims would require integrating knowledge that comes from more than one scientific article. While our retrieval-augmented-generation system presents the TOP-K paper snippets to the model making the feasibility assessment, in our baseline system these snippets come from a single search query, and multiple search queries may be required to collect different types of evidence that, together, could be integrated to improve the feasibility assessment. An initial pilot system that we constructed that iteratively allowed up to 5 rounds of evidence collection (each with a different query) before making an assessment did not appear to improve feasibility performance, suggesting integrating this knowledge into improved feasibility task performance is non-trivial. Similarly, it is plausible that building a specialized domain-specific corpus of supporting scientific articles may improve the utility of the retrieved knowledge, and increase task performance.

Limitations in Baseline Models (Code): Due to the current high estimated cost in running code generation systems, and the large size of the MATTER-OF-FACT test set, our code generation baseline was run in a highly limited form (i.e. short runtime, no debug iterations) that is best considered an approximate measure of *zero-shot code generation performance* on this feasibility assessment task, without the benefits of debugging iterations or the long experiment runtimes that code-based solutions to this task would almost certainly require. Some of these limitations are pragmatic, like cost and runtime, and are rapidly reduced as (for example) open language models for code generation that can be run on local hardware begin to approach paid API-BASED model performance. However, we ultimately believe that progress in code-based experimentation will require building systems that integrate materials science domain tools, software, and databases. We did not include any materials-science specific tooling in CODESCIENTIST's code retrieval library and required it to instead rely on the base model weights to implement these complex tool interfaces. It is highly likely that near-term improvement on this task will require a degree of manually integrating this tooling, just as other scientific agents (such as BIOME (Huang et al., 2025) in the biomedical domain) are currently using hand-build interfaces to domain-specific tools and databases to support their discovery tasks.

6 Conclusion

We present MATTER-OF-FACT, a benchmark for assessing the feasibility of 8.4K scientific claims in four high-impact subdomains of materials science: superconductors, semi-conductors, batteries, and aerospace materials. We frame the feasibility assessment task as a temporally-filtered claim verification task, and empirically demonstrate that strong baseline models using a variety of solving methods (including literature search and code generation) reach only modest performance on this task (72%). Performance on feasibility assessment can directly translate to improving automated scientific discovery systems, particularly in hypothesis generation, where filtering infeasible hypotheses can make scientific discovery more efficient, and lower overall experiment costs. Ultimate solution methods for the feasibility assessment task are likely to require a combination of reasoning over deep literature search, code-based simulation, and world modeling at the scale of subdomains.

Limitations

Temporal Filtering for Prediction: Temporal datasets that use the notion of backtesting offer the opportunity to construct prediction datasets for high-impact domains (e.g. Luo et al., 2018, link prediction for cancer biology) where the knowledge a system is predicting is potentially beyond current human knowledge, and for which gold labels are infeasible to construct. Temporal filtering assumes well-controlled models that have not been contaminated with data past their temporal filtering date. In this work we characterize the contamination rate of our base language models, and this analysis suggests that data contamination either does not exist, or is not a significant factor driving current performance. That being said, users of this benchmark should characterize the performance of novel base models to characterize how much data contamination may play a role.

Cost-Benefit Analyses: Pragmatically, to be useful for filtering scientific hypotheses, feasibility assessment methods must be able to perform well at scale. This necessitates that any pilot experiments (including code-based simulations) must be fast and inexpensive to run, otherwise the feasibility assessment step may be impractically expensive to provide overall cost savings. That being said, different applications and end-users may have varying preferred cost/performance points, and we encourage reporting performance as a function of overall cost (as we have done in this work) to help accurately assess the cost vs benefit of proposed models. It is our hope that providing a large-scale benchmark that necessitates developing inexpensive feasibility assessment methods will help facilitate innovation in this direction.

Acknowledgments

This research was developed with funding from the Defense Advanced Research Projects Agency's (DARPA) SciFy program (Agreement No. HR00112520300) to PJ at the University of Arizona. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. PJ has an outside interest in the Allen Institute for Artificial Intelligence. This interest has been disclosed to the University of Arizona and reviewed in accordance with its conflict of interest policies. We thank the members of the DARPA Scientific Feasibility (SciFy) program for thoughtful discussions. We also wish to thank the anonymous reviewers for their helpful comments, and their particular observation that literature-derived claims may be more polished than those present in earlier stages of the discovery process.

References

Jinheon Baek, Soyeong Jeong, Minki Kang, Jong Park, and Sung Hwang. 2023. Knowledge-augmented language model verification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1720–1736, Singapore. Association for Computational Linguistics.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chi Chen, Dan Thien Nguyen, Shannon J Lee, Nathan Baker, Ajay S Karakoti, Linda Lauw, Craig Owen, Karl T. Mueller, Brian A. Bilodeau, Vijayakumar Murugesan, and Matthias Troyer. 2024. Accelerating computational materials discovery with machine learning and cloud high-performance computing: from large-scale screening to experimental validation. *Journal of the American Chemical Society*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Preetam Prabhu Srikar Dammu, Himanshu Naidu, Mouly Dewan, YoungMin Kim, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. ClaimVer: Explainable claim-level verification and evidence attribution of text through knowledge graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13613–13627, Miami, Florida, USA. Association for Computational Linguistics.

Aghyad Deeb and Fabien Roger. 2024. Do unlearning methods remove information from language model weights? *arXiv preprint arXiv:2410.08827*.

Jiacheng Du, Zhibo Wang, Jie Zhang, Xiaoyi Pang, Jiahui Hu, and Kui Ren. 2024. Textual unlearning gives a false sense of unlearning. *arXiv preprint arXiv:2406.13348*.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715, Seattle, United States. Association for Computational Linguistics.

Ching Nam Hang, Pei-Duo Yu, and Chee Wei Tan. 2024. Trumorgpt: Query optimization and semantic reasoning over networks for automated fact-checking. In *2024 58th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, and 1 others. 2025. Biomni: A general-purpose biomedical ai agent. *biorxiv*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane A. Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24:251:1–251:43.

Peter Jansen, Oyvind Tafjord, Marissa Radensky, Pao Siangliulue, Tom Hope, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Daniel S Weld, and Peter Clark. 2025. CodeScientist: End-to-end semi-automated scientific discovery with code-based experimentation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13370–13467, Vienna, Austria. Association for Computational Linguistics.

John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andy Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, and 15 others. 2021. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, and 1 others. 2023. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Ruochen Li, Teerth Patel, Qingyun Wang, and Xinya Du. 2024. Mlr-copilot: Autonomous machine learning research based on large language models agents. *arXiv preprint arXiv:2408.14033*.

Ryan Liu and Nihar B Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, and 1 others. 2025. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pages 1–14.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7787–7813, Singapore. Association for Computational Linguistics.

Fan Luo, Marco A. Valenzuela-Escárcega, Gus Hahn-Powell, and Mihai Surdeanu. 2018. Scientific discovery as link prediction in influence and citation graphs. In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, pages 1–6, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. 2024. Eight methods to evaluate robust unlearning in llms. *arXiv preprint arXiv:2402.16835*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. 2025. Discoverybench: Towards data-driven discovery with large language models. In *The Thirteenth International Conference on Learning Representations*.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. 2023. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85.

Charles O'Neill, Tirthankar Ghosal, Roberta Răileanu, Mike Walmsley, Thang Bui, Kevin Schawinski, and Ioana Ciucă. 2025. Sparks of science: Hypothesis generation using structured paper data. *arXiv preprint arXiv:2504.12976*.

Raúl Ortega and José Manuel Gómez-Pérez. 2025. Sciclaims: An end-to-end generative system for biomedical claim analysis. *Preprint*, arXiv:2503.18526.

Jiefu Ou, William Gantt Walden, Kate Sanders, Zhengping Jiang, Kaiser Sun, Jeffrey Cheng, William Jurayj, Miriam Wanner, Shaobo Liang, Candice Morgan, Seunghoon Han, Weiqi Wang, Chandler May, Hannah Recknor, Daniel Khashabi, and Benjamin Van Durme. 2025. Claimcheck: How grounded are llm critiques of scientific papers? *ArXiv*, abs/2503.21717.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S Weld. 2024. Scideator: Human-llm scientific idea generation grounded in research-paper facet recombination. *arXiv preprint arXiv:2409.14634*.

Anku Rani, S.M Towhidul Islam Tonmoy, Dwip Dalal, Shreya Gautam, Megha Chakraborty, Aman Chadha, Amit Sheth, and Amitava Das. 2023. FACTIFY-5WQA: 5W aspect-based fact verification through

question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10421–10440, Toronto, Canada. Association for Computational Linguistics.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.

Jonathan Schmidt, Noah Hoffmann, Hai-Chen Wang, Pedro Borlido, Pedro J. M. A. Carriço, Tiago F. T. Cerqueira, Silvana Botti, and Miguel A. L. Marques. 2023. Machine-learning-assisted determination of the global zero-temperature phase diagram of materials. *Advanced Materials*, 35.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli. 2023. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6288–6304, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Zhiwei Zhang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. 2021. Abstract, rationale, stance: A joint model for scientific claim verification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3580–3586, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351.

A Utility for Hypothesis Filtering

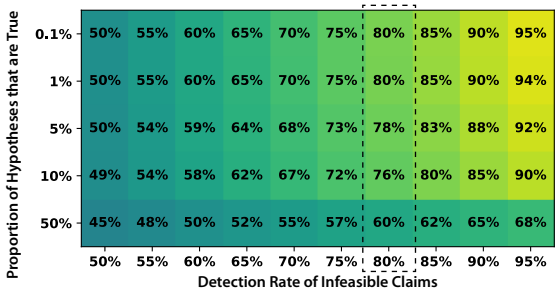

Figure 3: Relative efficiency (in terms of reduction in the number of experiments needed to be run) for a hypothetical automated scientific discovery (ASD) system that generates hypotheses with a certain true positive rate *(y-axis)*, after those hypotheses have been pre-filtered by a feasibility assessment system such as the models described in this work. This plot assumes a true positive (i.e. *feasible*) detection rate of 0.60, corresponding to the RAG (SEMANTICSCHOLAR, O4-MINI) model in Table 2, while the highlighted region corresponds to that model's infeasible claim detection rate (82%). For a hypothetical ASD system where 1% of the hypotheses it generated were *true/feasible*, the RAG model would reduce the number of experiments (i.e. cost) by 80%, while still discovering 60% of the true hypotheses.

Feasibility assessment has utility in impactful tasks such as (semi-automated) scientific discovery, particularly in the context of hypothesis generation. Hypothesis generation systems (e.g Lu et al., 2024; Jansen et al., 2025; O'Neill et al., 2025) have the capacity to generate an impractically large number of possible hypotheses (framed as claims) that one could test, and as a result running all their proposed experiments is costly (at best) and intractable (at worst). Coupling hypothesis generation with feasibility assessment would allow filtering out hy-

potheses that are unlikely to be feasible – i.e. yield experimental results that support the hypothesis – and ultimately increase the efficiency of scientific discovery systems in terms of the number of positive discoveries that can be made on a given budget. In automated hypothesis generation where overall likelihood of a hypothesis yielding positive results is low, increasing efficiency is dominated by correctly identifying (and filtering) infeasible hypotheses/claims. Figure 3 shows a plot of experiment efficiency (in terms of the reduction in the number of experiments that would need to be run) for hypothetical hypothesis generation systems that have different rates of generating true hypotheses, with the performance of the best-performing model (RAG (SEMANTIC SCHOLAR) using O4-MINI) highlighted. For a hypothetical hypothesis generation system where 1% of its hypotheses are true, this model would reduce the number of experiments needed to be run by approximately 80%, while still discovering 60% of the true hypotheses. This highlights that even systems with middle performance can have practical utility (in terms of cost savings) when coupled with scientific discovery systems.