NSF-SciFy: Mining the NSF Awards Database for Scientific Claims

Delip Rao[*†], **Weiqiu You**[†], **Eric Wong, Chris Callison-Burch**
University of Pennsylvania
Philadelphia, PA, USA
{delip, weiqiuy, exwong, ccb}@seas.upenn.edu

Abstract

We introduce NSF-SciFy, a comprehensive dataset of scientific claims and investigation proposals extracted from National Science Foundation award abstracts. While previous scientific claim verification datasets have been limited in size and scope, NSF-SciFy represents a significant advance with 2.8 million claims from 400,000 abstracts spanning all science and mathematics disciplines. We present two focused subsets: NSF-SciFy-MatSci with 114,000 claims from materials science awards, and NSF-SciFy-20K with 135,000 claims across five NSF directorates. Using zero-shot prompting, we develop a scalable approach for joint extraction of scientific claims and investigation proposals. We demonstrate the dataset's utility through three downstream tasks: non-technical abstract generation, claim extraction, and investigation proposal extraction. Fine-tuning language models on our dataset yields substantial improvements, with relative gains often exceeding 100%, particularly for claim and proposal extraction tasks. Our error analysis reveals that extracted claims exhibit high precision but lower recall, suggesting opportunities for further methodological refinement. NSF-SciFy enables new research directions in large-scale claim verification, scientific discovery tracking, and meta-scientific analysis[1].

1 Introduction

The overall growth rate of scientific publications is estimated to be 4% annually, with a doubling time of 17 years (Bornmann et al., 2021). Within this deluge, researchers, reviewers, and the general public struggle to separate substantiated claims from spurious ones—whether it is the "quantum supremacy" assertions in computing, the short-lived excitement over LK-99 superconductors[3], or the misunderstanding surrounding microplastic leaches from black plastic spatulas[4]. Manual verification of ever growing body of scientific claims has become intractable, yet the economic and societal consequences of unverified claims are increasingly severe.

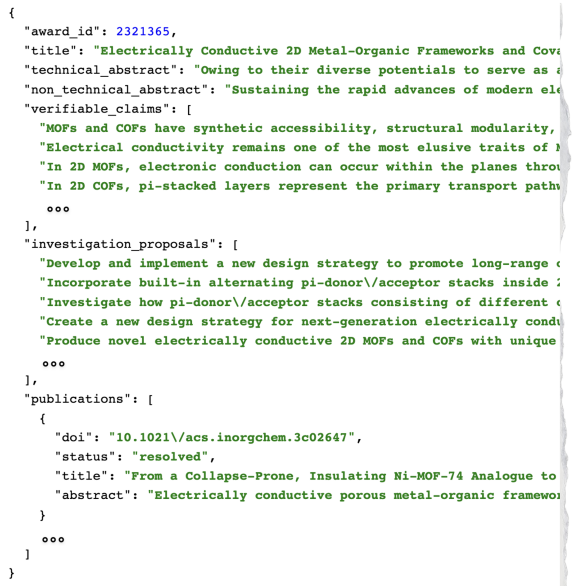

Figure 1: A sample record from our dataset. Each record contains 1) Award ID and title, 2) NSF Directorate, 3) Technical and non-technical abstracts, 4) Scientific Claims, 5) Investigation Proposals, and 6) Associated publications, when present.

Wadden et al. (2020) introduced the task of scientific claim verification with the SciFACT dataset, focusing primarily on automatic verification of scientific claims. Follow up works (see Section 2 for a detailed account) have mostly focused on the healthcare, building datasets from scientific publications, and modest-sized dataset creation. In

[*]Corresponding author, [†]co-first author

[1]Code and data available at https://github.com/darpa-scify/NSFSciFy

[3]for an entertaining digression c.f., https://en.wikipedia.org/wiki/LK-99

[4]c.f., https://nationalpost.com/news/canada/black-plastic

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 45679–45697

Dataset	# claims	# docs	Evidence Source	Domain
SciFACT (Wadden et al., 2020)	1.4K	5K	Research papers	Biomedical
PubHEALTH (Kotonya and Toni, 2020)	11.8K	11.8K	Fact-checking sites	Public health
CLIMATE-FEVER (Diggelmann et al., 2020)	1.5K	7.5K	Wikipedia articles	Climate change
HealthVer (Sarrouti et al., 2021)	1.8K	738	Research papers	Healthcare
COVID-Fact (Saakyan et al., 2021)	4K	4K	Research, news	COVID
CoVERT (Mohr et al., 2022)	300	300	Research, news	Biomedical
SciFACT-Open (Wadden et al., 2022)	279	500K	Research papers	Biomedical
NSF-SciFy-MatSci (ours)	**114K**	**16K**	**NSF award abstracts**	**Material Science**
NSF-SciFy-20K (ours)	**135K**	**20K**	**NSF award abstracts**	**All Science & Math**
NSF-SciFy (ours)	**2.8M**	**400K**	**NSF award abstracts**	**All Science & Math**

Table 1: **(NSF-SciFy spans all science and math *domains* and includes diverse *data types*: technical/non-technical abstracts, claims, and investigation proposals.)** While previous datasets like SciFACT and PubHEALTH contain at most thousands of claims from published research papers or fact-checking sources, our NSF-SciFy-MatSci and NSF-SciFy-20K datasets individually contribute more than 100K claims. The full NSF-SciFy dataset represents an order-of-magnitude increase with 2.8M claims across 400K abstracts spanning all science & math disciplines. This work introduces grant abstracts as a novel, untapped source for scientific claim extraction, complementing existing approaches that focus on published literature, news articles, or social media.

this work, we relax all of these aspects and look at building at least an order of magnitude large-scale scientific claim dataset covering all of basic science. We envision building of such large-scale, scientific claim datasets to help future work on robust scientific claim verification systems.

We introduce NSF-SciFy[2], a comprehensive dataset of claims and investigation proposals extracted from National Science Foundation (NSF) award abstracts. We choose NSF abstracts as our source material for several reasons:

1. NSF is a primary driver of U.S. scientific innovation, funding approximately 25% of all federally supported basic research, spanning the entirety of science and math areas, with an annual budget of $9.9 billion (FY 2023). Any claim dataset derived from the NSF awards database should faithfully represent the scientific Zeitgeist.
2. NSF's rigorous subject matter expert-review process provides a high-quality filter for the claims made in funded proposals.
3. The public availability and permissive usage terms of the NSF awards database makes it an excellent resource for open science research.
4. Previous datasets on scientific claims have been derived from scientific papers, but claims in scientific grants, and particularly investigation proposals, remain unstudied.

While not the focus of this paper, grant award abstracts additionally provide a unique opportunity to study the relationship between what researchers claim and what they propose to investigate. This could offer valuable insights into scientific practice and the evolution of research questions.

In this paper, we make the following contributions: (1) We introduce NSF-SciFy, the largest scientific claim dataset to date with 2.8M claims extracted from 400K NSF award abstracts, establishing grant proposals as a novel source for scientific claim extraction; (2) We create NSF-SciFy-MatSci focusing exclusively on materials science with 114K extracted claims from 16K abstracts. This is the first materials science claim dataset and, in number of extracted claims, this alone is an order of magnitude bigger than the largest publicly available claim dataset; In addition, we also create NSF-SciFy-20K with 135K claims spanning five NSF directorates. (3) We develop a zero-shot prompting approach for joint extraction of scientific claims and investigation proposals as a scalable way to bootstrap high-precision, large-scale scientific claim datasets; (4) We present novel evaluation metrics for claim/proposal extraction based on LLM judgments, showing that fine-tuned models significantly outperform base models; and (5) Finally, we release all datasets and trained models from our work for unfettered research and commercial use. Our dataset and methods enable new opportunities for large-scale claim verification, scientific discovery tracking, and meta-scientific research. See Appendix A for reproducibility statement.

[2]Short for "NSF SCIentific FeasibilitY".

2 Related Work

Scientific claim extraction and verification has emerged as an important research area as the volume of scientific literature continues to grow exponentially. Previous work has primarily focused on claims from published papers, fact-checking sites, and news articles.

Scientific Claim Datasets Several datasets have been developed for scientific claim verification, but all have focused on claims from published literature, while we undertake the study of grant award abstracts. SciFACT (Wadden et al., 2020) contains 1,400 scientific claims derived from research papers in the biomedical domain. PubHEALTH (Kotonya and Toni, 2020) includes 11,800 claims from journalists and fact-checkers in public health. CLIMATE-FEVER (Diggelmann et al., 2020) compiled 1,500 claims from news articles about climate change. HealthVer (Sarrouti et al., 2021) extracted 1,800 claims from search queries related to health topics. COVID-Fact (Saakyan et al., 2021) and CoVERT (Mohr et al., 2022) focused on COVID-19 related claims from social media. SciFact-Open (Wadden et al., 2022) expanded the original SciFact dataset using information retrieval pooling, yet it still remains health-care focused and a few orders of magnitude smaller than our largest dataset.

Table 1 situates existing scientific claim datasets with our NSF-SciFy datasets, highlighting the significantly larger scale of our contribution (2.8 million claims in NSF-SciFy, 135,000 claims inNSF-SciFy-20K and 114,000 claims in NSF-SciFy-MatSci), broad topic coverage (all of science and math), and novelty of data source (grant abstracts). See Figure 2.

Meta Science and Social Science Previous works have examined grants data in social science and meta-science contexts. For example, Park et al. (2024) examine the relationship between interdisciplinary grants and the impact of papers they support and Xu et al. (2022) study the influence of research funding on team structure using grant data. While these are tenuously connected to our work, we list them for the sake of completeness.

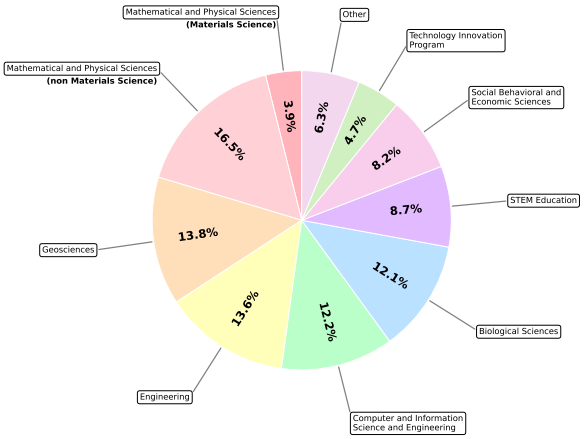

Figure 2: **(NSF-SciFy contains a large variety of domains.)** Distribution of awards areas as represented by the National Science Foundation directorates in NSF-SciFy, illustrating the breadth and comprehensiveness of scientific claims in our dataset. The NSF-SciFy-MatSci subset spanning all of materials science awards represents 3.9% of the entire dataset.

3 Building NSF-SciFy

3.1 Data Collection

We downloaded the entire NSF Awards database[3] in XML format, containing more than 0.5 million awards from 1970 through September 2024. After parsing, we obtained 412,155 parseable awards, which we call NSF-SciFy.

In this paper, we focus on all awards from the Division of Materials Research (DMR), which is responsible for most materials science awards at the NSF. This subset, called NSF-SciFy-MatSci, contains 16,031 awards, representing approximately 3.2% of the entire NSF awards database. We chose materials science as our focus due to its interdisciplinary nature and technological importance. In addition, we build NSF-SciFy-20K, a different subset of 20K awards spanning 5 NSF directorates — Mathematical and Physical Sciences (MPS), Geological Sciences (GEO), Engineering (ENG), Computer and Information Science and Engineering (CSE), and Biological Sciences (BIO).

3.2 Data Processing

As Figure 1 illustrates, each record in NSF-SciFy-MatSci typically contains:

1. Award ID, title, and year.
2. Directorate and division information
3. Technical abstract

[3] https://www.nsf.gov/awardsearch/advancedSearch.jsp

4. Non-technical abstract (present in ∼81% of awards)
5. Scientific claims made in the abstracts
6. Investigation proposals in the abstracts
7. Publications resulting from the grant (when available)

The practice of updating awards with resulting publications is relatively recent, primarily occurring from 2014 onwards. For awards where publications are present, we extracted the DOIs and resolved them to obtain titles, abstracts, and publication URLs.

3.3 Claim and Investigation Proposal Extraction

To extract scientific claims and investigation proposals from the award abstracts, we developed a zero-shot prompting approach using Anthropic's Claude-3.5[4] model. Our prompt instructed the model to identify two types of statements:

1. **Claims**: Statements that the abstract claims to be true or states as assumptions, either explicitly or implicitly. [5]
2. **Investigation proposals**: Forward-looking statements that propose specific research activities as part of the award.

We structured the prompt to return a JSON object containing the award ID, technical abstract, non-technical abstract, a list of claims, and a list of investigation proposals. To maintain consistency and quality, we set temperature to zero for all extractions. See Appendix B for the exact prompt and Appendix G for sample claims and investigation proposals.

We performed qualitative experiments with several prompt variants and our analysis showed that jointly extracting claims and investigation proposals helped maintain the relevance of extracted claims. When claims were extracted without also extracting investigation proposals, the model often confused forward-looking statements about proposed investigations as factual claims.

4 Dataset Analysis

NSF-SciFy The full dataset contains 412,155 award abstracts spanning from 1970 to 2024, with 2.8 million scientific claims and corresponding investigation proposals.

NSF-SciFy-MatSci This materials science subset, which is the focus of this preprint, contains:

- 16,042 awards with each with a technical and non-technical abstract
- 114K extracted scientific claims (average of 7 ± 2 claims per abstract-pair)
- 145K extracted investigation proposals (average of 9 ± 3 proposals per abstract-pair)
- 2,953 awards with linked publications (18.4% of the dataset). Such awards had anywhere between 1 – 4 publications.

NSF-SciFy-20K For building models across all NSF directorates, we take 20,000 sample subset of NSF-SciFy, by stratifying across 5 directorates.

- 20,001 awards with each with a technical and non-technical abstract
- 135K extracted scientific claims (average of 7 ± 2 claims per abstract-pair)
- 139K extracted investigation proposals (average of 7 ± 2 proposals per abstract-pair)

4.1 Technical vs. Non-Technical Abstracts

We investigated the differences between technical and non-technical abstracts in our dataset. Using a symmetric BLEU score to measure textual similarity between paired abstracts, we found that only 202 (1.5%) out of 13,025 technical/non-technical abstract pairs had a similarity score greater than 0.6, suggesting that the non-technical abstracts are not simply copied from the technical abstracts.

Since grant abstracts are previously unexamined in literature, we further investigated the stylistic differences between technical and non-technical abstracts using pre-trained document embedding models. Figure A7 compares content embeddings from SPECTER (Cohan et al., 2020) and style embeddings from STEL (Patel et al., 2025). Using these embeddings with a linear SVM classifier, we achieved F1 scores of 90.99 (SPECTER), 88.42 (STEL), and 89.99 (concatenated), demonstrating that the abstracts are distinguishable both in content and style.

4.2 Taxonomies of Claims and Investigation Proposals

Claims. To characterize the types of assertions made in NSF award abstracts, we analyzed 810 extracted claims from 120 awards sampled across five NSF directorates (MPS, GEO, ENG, CSE, BIO). We identified eight broad categories, covering well-known facts, observed phenomena, ap-

[4] `Claude-3.5-Sonnet-20240620` accessed between Sep-Oct. 2024, to be specific.

[5] Our notion of claims follows prior work (Tang et al., 2024).

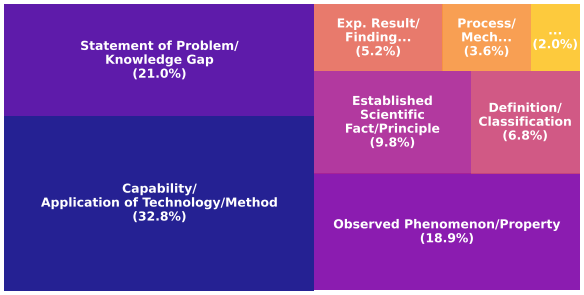

Figure 3: **(Most scientific claims in the abstracts are about knowledge gap and application methods.)** A treemap of the scientific claim categories in NSF awards. See Table A10 for descriptions of these categories.

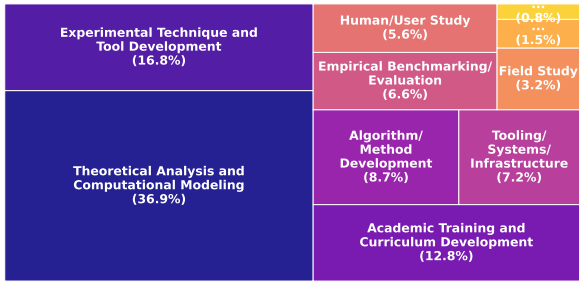

Figure 4: **(Most investigation proposals in the abstracts are about experimental technique and theoretical analysis.)** A treemap of the investigation proposal categories in NSF awards. See Table A11 for descriptions of these categories.

plications of methods or technologies, theoretical predictions, experimental findings, knowledge gaps, definitions/classifications, and process descriptions. Figure 3 shows their distribution. The most common types are *Capability/Application of Technology/Method* (32.8%), *Statement of Problem/Knowledge Gap* (21.0%), and *Observed Phenomenon/Property* (18.9%). Examples for all categories are shown in Table A10.

Investigation Proposals. We performed a parallel analysis on 833 investigation proposals from the same award set, identifying eight categories spanning theoretical analysis, experimental technique development, algorithm/method development, academic training, and various empirical study types. Figure 4 shows their distribution. The majority fall under *Theoretical Analysis and Computational Modeling* (36.9%), *Experimental Technique and Tool Development* (16.8%), and *Academic Training and Curriculum Development* (12.8%). Examples for all categories are shown in Table A11.

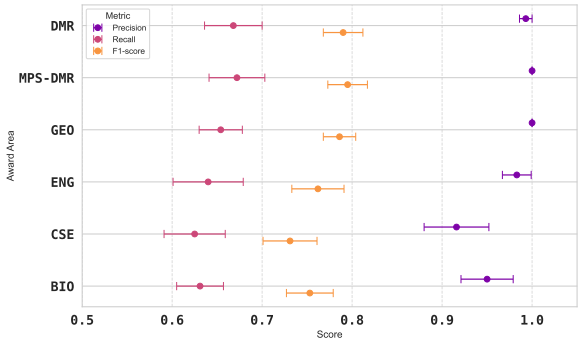

Figure 5: **(Claim extraction achieves consistently high precision across all areas, while recall is lower, leading to moderate F1-scores.)** A Cleveland dot plot of precision, recall, and F1-score across different NSF Award Areas for claims extracted via Claude (See Section 3.3). Error bars denote standard deviation (bootstrap N=1000). See Section 4.3 for analysis.

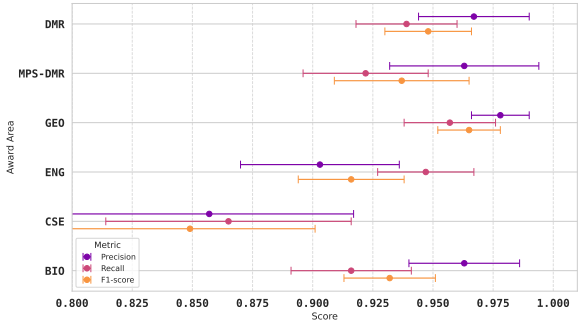

Figure 6: **(Investigation Proposal extraction achieves consistently high precision across all areas, while recall is lower, leading to moderate F1-scores.)** A Cleveland dot plot of precision, recall, and F1-score across different NSF Award Areas for investigation proposals extracted via Claude (See Section 3.3). Error bars denote standard deviation (bootstrap N=1000). See Section 4.3 for analysis.

4.3 Evaluating Extracted Claims and Investigation Proposals

We evaluate the quality of the extracted claims and investigation proposals (Section 3.3) by manually annotating 120 sampled awards (Section 4.2) and computing precision, recall, and F1. For each of the six NSF areas—Materials Science (DMR), Mathematical and Physical Sciences excluding Materials Science (MPS-DMR), Geological Sciences (GEO), Engineering (ENG), Computer and Information Science and Engineering (CSE), and Biological Sciences (BIO)—we randomly sampled 20 items per area. Using `GPT-4o` (OpenAI, 2024), we identified additional true elements G' missed by the extracted set (with $FN = |G'|$) and categorized

previously extracted elements as correct (TP) or incorrect (FP). Annotators (PhD students) manually verified `GPT-4o`'s outputs on 20 abstracts and confirmed near-perfect verification accuracy. Precision, recall, and F1 were then computed using FN, TP, and FP.

Figures 5 and 6 summarize performance across the six areas for claims and investigation proposals, respectively. For claims, extraction achieves consistently high precision but lower recall, leading to moderate F1-scores. For investigation proposals, precision, recall, and F1 are more balanced across areas, indicating more comprehensive coverage. Overall, the extracted data is of high quality, though improving recall for claims remains an important direction.

5 Tasks, Metrics, and Experiments

Previously, Section 3.3 describes the data extraction process using a large model, and Section 4 evaluates the quality of the resulting synthetic data. Here, we demonstrate its utility by evaluating the performance of smaller models fine-tuned on it across three NLP tasks:

1. The **Non-technical Abstract Generation** task translates dense, technical grant abstracts into accessible language for broader science communication. Motivated by capturing the core scientific essence while navigating stylistic and content differences between technical and lay summaries, this task uses the dataset's paired examples (common in NSF awards) to train models for this nuanced transformation.
2. The **Abstract to Scientific Claims Extraction** task automates identifying verifiable assertions—the core of scientific discourse—from grant abstracts, which capture these claims at an early, pre-publication stage. Significant performance gains post-fine-tuning highlight the dataset's effectiveness in teaching models to pinpoint these crucial statements.
3. The **Abstract to Investigation Proposals Extraction** task distinguishes aspirational research intentions from established claims, offering a novel analysis of scientific texts. This provides a clearer view of the planned research trajectory by identifying intended activities. It complements claim extraction by presenting a fuller picture of proposed work, from assertions to investigative pathways, again showing significant fine-tuning efficacy due to the dataset's focused nature.

To explore the three tasks, we finetuned two 7B parameter language models:

- `Mistral-7B-instruct-v0.3` (Jiang et al., 2023)
- `Qwen2.5-7B-Instruct` (Yang et al., 2024)

5.1 Data Preparation

Starting with 16,042 processed entries in NSF-SciFy-MatSci, we removed near-duplicates in technical and non-technical abstracts using trigram Jaccard similarity (threshold > 0.9), resulting in 11,569 data points. We further filtered cases where character-level 10-gram similarity between an entry's technical and non-technical abstracts exceeded 0.6, yielding 11,141 final data points. We split this dataset into train/validation/test sets with 8,641/500/2,000 examples, respectively.

5.2 Finetuning Details

For fine-tuning, we used LoRA (Hu et al., 2021) with rank=128, lora_alpha=64 and a learning rate of 1e-5 scheduled linearly. We updated the query, key, value, and output projection layers, as well as MLP gate, up, and down projections. We ran the finetuning on an A100 GPU for 3 epochs, 100 warmup steps, and a batch size of 2 with 4 accumulated steps. Each epoch takes around one hour.

5.3 Evaluation Metrics

For Task 1 – abstract generation – we employed a comprehensive evaluation framework using both BERTScore (Zhang* et al., 2020) and ROUGE (Lin, 2004) metrics to assess the quality of generated non-technical abstracts. This combination enables us to capture both lexical overlap and structural similarity through the ROUGE variants, while BERTScore provides insights into semantic alignment between the generated texts and reference abstracts. Incorporating such multi-viewed metrics[6] ensures that the evaluation reflects not only the presence of key words and phrases but also the underlying meaning and narrative coherence of the abstracts.

For Task 2 – claim extraction – we developed a novel evaluation approach using LLM-based comparisons. Previous methods for claim evaluations focused on comparing a single claim against a single document. See Tang et al. (2024), for example.

[6]For BERTScore we report precision, recall and F1, and for ROUGE we report ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-L-sum.

However, our setting required evaluating a set of extracted claims against a gold set of claims.

Towards that end, we defined a boolean function $\mathbf{\Phi}_{\text{claim}}$ using `GPT-4o-mini` (OpenAI, 2024) with zero-shot prompting to determine whether a generated claim is supported by a gold standard claim. See Appendix C for prompt details[7]. Using this function, we calculated precision and recall as follows:

$$\text{Precision} = \frac{1}{|S|}\sum_{c\in S}\max_{g\in G}\mathbf{\Phi}_{\text{claim}}(c,g)$$

$$\text{Recall} = \frac{1}{|G|}\sum_{g\in G}\max_{c\in S}\mathbf{\Phi}_{\text{claim}}(g,c)$$

where S is the set of claims generated from the finetuned model, after removal of any repeats/near-repeats [8], and G is the gold standard set. We note that this is a variant of precision/recall metrics defined for image captioning in (Deitke et al., 2024), however unlike Deitke et al., we explicitly use $\mathbf{\Phi}_{\text{claim}}$ in computing both precision and recall. This is necessary as we need to accurately penalize any spurious claims generated by the finetuned model. Works by (Gu et al., 2025; Liu et al., 2023) are relevant here.

We carefully validated our LLM on a subset of 120 awards using human annotators assisted by `GPT-4o-mini`. We restricted the role of GPT-4o-mini to only pairwise sentence comparison, a task which prior work has shown as easy for large foundation models. We found a near-perfect correlation between human judgments and `GPT-4o-mini`'s judgements for this pairwise comparison [9]. Based on this validation, we applied LLM-as-judge evaluation to the full dataset, a scale that would otherwise have been infeasible to annotate manually. All P/R/F1 values were computed deterministically using the pairwise outputs.

Analogously, for Task 3 – extraction of investigation proposals – we define precision and recall similarly but use a different pairwise boolean judge function $\mathbf{\Phi}_{\text{IP}}$ *mutatis mutandis*. See Appendix D for prompt details.

[7]We tried several slight edits of the prompts and found them to be robust to such changes.

[8]We determine repeats and near-repeats in the generation by thresholding cosine similarity calculated over a TF-IDF representation of the generated claims.

[9]We use `GPT-4o-mini` here because this is a simple task and we found `GPT-4o-mini` sufficient.

Metric	Mistral	Qwen
BERTScore-P	0.8563 (+0.38% ↑)	0.8459 (+0.98% ↑)
BERTScore-R	0.8555 (+0.30% ↑)	0.8597 (+1.61% ↑)
BERTScore-F1	**0.8561** (+0.36% ↑)	0.8437 (+0.75% ↑)
ROUGE1	0.2000 (+2.58% ↑)	0.1978 (+1.98% ↑)
ROUGE2	0.0198 (+4.76% ↑)	0.0210 (+3.89% ↑)
ROUGE-L	0.1273 (+2.96% ↑)	0.1466 (+0.65% ↑)
ROUGE-L-sum	0.2166 (+2.45% ↑)	0.2078 (+1.66% ↑)

Table 2: **(Finetuned models have modest improvements on technical abstract to non-technical abstract translation, indicating excellent out-of-the-box performance for this task.)** Finetuning performance for `Mistral-7B-instruct-v0.3` and `Qwen2.5-7B-Instruct` models for Technical abstract to Non-technical abstract translation (Task 1), with relative improvements over the corresponding unfinetuned model indicated in green. Error bars for all metrics at 95% confidence intervals range between 0.0000–0.0025. Mistral model outperforms Qwen on almost all metrics for this task regardless of finetuning.

Metric	Mistral	Qwen
Precision	0.7450 (+116.7% ↑)	0.6839 (+107.1% ↑)
Recall	0.7098 (+59.5% ↑)	0.6611 (+7.8% ↑)
F1	0.7097 (+101.8% ↑)	0.6541 (+63.3% ↑)

Table 3: **(Finetuning leads to large improvements in claim extraction from abstracts.)** Finetuning performance for `Mistral-7B-instruct-v0.3` and `Qwen2.5-7B-Instruct` models for Claim Extraction from abstracts (Task 2), with relative improvements over the corresponding unfinetuned model indicated in green. Error bars for all metrics at 95% confidence intervals range between 0.0038–0.0055. Mistral model outperforms Qwen on almost all metrics for this task regardless of finetuning. We note the large positive percent changes, sometimes improvements as large as 2x, indicate finetuning is indispensable for claim extraction. Mistral model outperforms Qwen on almost all metrics for this task.

6 Results

6.1 Non-technical Abstract Generation

Table 2 shows the results for Task 1. Both Mistral and Qwen models demonstrated strong performance, with fine-tuning providing modest improvements. The Mistral model outperformed Qwen on almost all metrics, achieving a BERTScore-F1 of 0.8561 after fine-tuning (+0.36% relative improvement). ROUGE scores were generally low (0.01-0.22), reflecting the stylistic differences between technical and non-technical abstracts.

6.2 Scientific Claim Extraction

For Task 2 (claim extraction), fine-tuning yielded substantial improvements. As shown in Table 3, the fine-tuned Mistral model achieved a precision of 0.7450 (+116.7% relative improvement), recall of 0.7098 (+59.5%), and F1 of 0.7097 (+101.8%). The Mistral model consistently outperformed Qwen, though both showed significant benefits from fine-tuning.

6.3 Investigation Proposal Extraction

Similarly, Task 3 (proposal extraction) showed dramatic improvements with fine-tuning. As shown in Table 4, the Mistral model achieved a precision of 0.7351 (+18.24%), recall of 0.7539 (+127.24%), and F1 of 0.7261 (+90.97%) after fine-tuning. The relative improvements were even larger for the Qwen model, though Mistral still performed better overall.

Since Mistral models seemed to have an edge over the Qwen2.5 models for these tasks, we also trained a Mistral only version of on the NSF-SCIFY-20K subset which spans all NSF directorates. The results can be found in Appendix F.

Metric	Mistral	Qwen
Precision	0.7351 (+18.24% ↑)	0.7245 (+70.07% ↑)
Recall	0.7539 (+127.24% ↑)	0.6865 (+81.57% ↑)
F1	0.7261 (+90.97% ↑)	0.6827 (+112.60% ↑)

Table 4: **(Finetuning leads to large improvements in investigation proposal extraction from abstracts.)** Finetuning performance for `Mistral-7B-instruct-v0.3` and `Qwen2.5-7B-Instruct` models for extraction of Investigation Proposals from award abstracts (Task 3), with relative improvements over the corresponding unfinetuned model indicated in green. Error bars for all metrics at 95% confidence intervals range between 0.0036–0.0073. Mistral model outperforms Qwen on almost all metrics for this task regardless of finetuning. We note the large positive percent changes, sometimes improvements as large as 2x, indicate finetuning is indispensable for this task. Mistral model outperforms Qwen on almost all metrics for this task.

7 Error Analysis

We conduct error analyses on both claim extraction and investigation proposal extraction to understand common failure modes of fine-tuned models.

Claims. Using 120 awards from the test sets of NSF-SCIFY-MATSCI and NSF-SCIFY-20K, we examined 802 claims generated by a fine-tuned Mistral-7B model and found an error rate of 2.6%. We categorized the errors into five types: (1) **Overconfidence** — misrepresenting hedged statements as factual assertions; (2) **Mixing Information** — combining content from multiple sentences incorrectly; (3) **Overgeneralization** — extending claims beyond what is stated; (4) **Information Omission** — dropping key qualifiers and altering meaning; and (5) **Administrative Hallucinations** — inserting funding or institutional information not present. Overconfidence and overgeneralization were the most common. Claude-extracted claims had a slightly lower error rate (2.1%), mostly administrative hallucinations.

Investigation Proposals. A parallel analysis on 833 investigation proposals yielded an error rate of 2.4%. We identified four error types: (1) **No Investigation Proposals** — generating proposals when none exist in the abstract; (2) **Content Mismatch** — introducing or omitting key elements; (3) **Overspecification** — adding unsupported details; and (4) **Existing Work** — describing prior work rather than forward-looking plans.

Examples per error type are in Appendix I. Mitigation strategies across both tasks include uncertainty calibration, and stricter alignment between extractions and source text. We manually check 20 examples and found most "correct" claims are indeed correct, while over half of the "errors" are not actual errors, suggesting even higher true accuracy.

8 Discussion and Conclusion

We introduced NSF-SCIFY, a large dataset of 2.8 million scientific claims and proposals from 400,000 NSF grant abstracts across all science and mathematics disciplines. Focused subsets include NSF-SCIFY-MATSCI (114,000 materials science claims) and NSF-SCIFY-20K (135,000 claims from five directorates). Experiments demonstrate that fine-tuning language models on NSF-SCIFY significantly improves scientific claim and proposal extraction, with relative performance gains often exceeding 100%. Non-technical abstract generation saw modest improvements due to strong baselines. Stylistic differences between technical and non-technical abstracts offer potential for science communication. Our claim taxonomy identifies prevalent assertion types like capability/application and problem/knowledge gap statements. NSF-SCIFY's unique advantages include its vast scale,

high quality from NSF expert review, comprehensive coverage of scientific domains, a temporal span from 1970-2024 enabling longitudinal studies, and, for recent grants, links to resulting publications. NSF-SCIFY opens new research avenues in large-scale claim verification, scientific discovery tracking, and meta-scientific analysis, a key resource for understanding scientific assertions at their origin.

Limitations

Source Material Scope. The dataset, derived from NSF award abstracts, offers insights into early-stage scientific claims from a rigorously reviewed, cross-disciplinary source. However, it currently excludes claims from unfunded proposals or international contexts. Future work may expand to other agencies and sources.

Bias and Coverage Considerations. While the dataset currently excludes unfunded and international proposals, the National Science Foundation accounts for approximately 25% of U.S. federally supported basic research, providing substantial coverage across scientific disciplines. We also note an availability bias: (1) unfunded proposals are not publicly accessible, aside from a handful of exemplars shared online, and (2) international proposals are rare and geographically dispersed. Given their importance, systematically incorporating international proposals represents an important direction for future work.

Extraction Methodology. Our approach utilizes zero-shot prompting with large language models, refined by prompt engineering and selective human validation. While manual evaluation shows consistently high precision across all directorates, our zero-shot extraction pipeline exhibits lower recall. At this bootstrapping stage, this was a deliberate design choice – we prioritized high precision to ensure the foundational reliability of the extracted statements and to prevent the proliferation of spurious claims. As demonstrated in Table 3, fine-tuning smaller models on this dataset significantly improves extraction, roughly doubling the F1 score for both claim and proposal tasks by significantly boosting recall alongside precision. Furthermore, the massive scale of NSF-SCIFY enables data-intensive strategies to close the recall gap in future work. For instance, the dataset's massive size and cross-disciplinary diversity provide the necessary training signals for multi-pass extraction protocols, allowing models to iteratively capture secondary claims. It also serves as a robust foundation for fine-tuning diverse open-source models for agreement-based ensembling. Finally, the vast candidate pool allows for targeted active annotation, enabling researchers to isolate and manually label only the most complex, low-confidence edge cases to systematically improve recall.

Evaluation Design. We introduced LLM-based metrics for evaluating claims and investigation proposals, offering a nuanced assessment beyond lexical overlap. These metrics correlate well with human judgment in samples, but broader validation across more scientific domains is needed to confirm their robustness. The public dataset and code aim to facilitate such community efforts.

Temporal and Linked Data Coverage. Spanning over five decades and including recent linked publication metadata, the dataset's systematic outcome tracking is limited for older awards. This restricts longitudinal analysis of claim evolution from proposal to publication. Broader, consistent outcome reporting could enrich NSF-SCIFY for deeper research trajectory studies.

Generalizability. While designed and validated for National Science Foundation abstracts, whose structure may differ from other scientific communications, the general framework is adaptable. It could be extended to related corpora like other funding agencies, patent abstracts, or scientific news, creating opportunities for future research.

Baselines. We report results using two competitive baseline models — Mistral-7B-v0.3 and Qwen2.5-7B — and observe consistent trends across both. We do not include additional baselines in this work; a more extensive comparison with other models is left for future work. All datasets and models are publicly released to facilitate such comparisons.

Acknowledgments

The authors would like to acknowledge NSF award CCF 2442421, the AI2050 program at Schmidt Sciences (Grant G-25-67983), the Defense Advanced Research Projects Agency (DARPA) SciFy program (Agreement No. HR00112520300) for funding this research, and the Office of the Director of National Intelligence (ODNI), Intelligence Ad-

vanced Research Projects Activity (IARPA), via 56000026C0019. We also thank the National Science Foundation (NSF) for making award data publicly available, enabling this research. Any views, opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the official policy, position, or views, either expressed or implied, of the National Science Foundation, DARPA, the Department of Defense, ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1).

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *Preprint*, arXiv:2409.17146.

Thomas Diggelmann, Jordan L. Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A dataset for verification of real-world climate claims. *CoRR*, abs/2012.00614.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Isabelle Mohr, Amelie Wührl, and Roman Klinger. 2022. CoVERT: A corpus of fact-checked biomedical COVID-19 tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 244–257, Marseille, France. European Language Resources Association.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed: 2025-10-01.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Minsu Park, Suman Kalyan Maity, Stefan Wuchty, and Dashun Wang. 2024. Interdisciplinary papers supported by disciplinary grants garner deep and broad scientific impact. *Preprint*, arXiv:2303.14732.

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2025. StyleDistance: Stronger content-independent style embeddings with synthetic parallel examples. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*,

pages 8662–8685, Albuquerque, New Mexico. 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.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. 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.

Fengli Xu, Lingfei Wu, and James A. Evans. 2022. Quantifying hierarchy in scientific teams. *Preprint*, arXiv:2210.05852.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Appendix

A Reproducibility Statement

To foster research on large-scale claim extraction, we are releasing our datasets, training code, and trained models:

- NSF-SciFy-MatSci: Materials Science subset with extracted claims, investigation proposals, and resolved publication information.
- NSF-SciFy: Similar in content to NSF-SciFy-MatSci, but a larger superset spanning all of NSF awards database. The key difference is the claims and investigation proposals are extracted from our finetuned models instead of frontier LLMs.
- Our code is available at `https://github.com/darpa-scify/NSFSciFy`.
- Our best finetuned model checkpoints for extraction of claims and investigation proposals at `https://huggingface.co/darpa-scify/nsf-scify-matsci-claims`.
- License: We will release our data and model under apache-2.0.
- We used all existing artifacts in accordance with their intended research purposes, and we specify that NSF-SciFy is released solely for research and commercial use under compatible access conditions.

B Complete Prompt for Extracting Claims and Investigation Proposals

You are an expert materials science researcher. Given an input JSON description of an NSF material science award abstract, parse out the technical and nontechnical abstracts, and identify the claims and research/investigation proposals the abstract makes. Be thorough. Answer in the following JSON format:

```
{
  "award_id": "", // copied from input
  "technical_abstract": "" // technical
      abstract if present, otherwise
      contents of the abstract field in the
       input
  "non_technical_abstract": /non-technical
      abstract if present, otherwise empty
  "claims": [ // list of strings
  ],
  "investigation_proposals": [ // list of
```

```
      strings
  ],
}
```

claims are statements that the abstract claims to be true or states as an assumption explicitly or implicitly.
investigation_proposals are forward-looking statements that the abstract proposals to investigate as a part of this award.
Ensure that the output is in JSON format and that the JSON is valid.

We manually tested the prompt with a few award abstracts to make sure it was optimal for this task.

C Prompt for Task 2 evaluation function Φ_{claim}

Check two scientific claims c1 and c2, if c1 is supported by c2. If c2 includes all the evidences for c1, but also includes additional content, then it should still be supported (YES). If not all information of c1 is included in c2, or if c2 contains information that conflicts with information in c1, then it should be unsupported (NO). Answer only as a YES or NO.
c1: {c1}
c2: {c2}

D Prompt for Task 3 evaluation function Φ_{IP}

Check two investigation proposals c1 and c2, if c1 is supported by c2. If c2 includes all the investigations proposed by c1, but also includes additional proposals, then it should still be supported (YES). If not all proposed investigations by c1 is included in c2, or if c2 contains investigation actions that conflict with investigation actions in c1, then it should be unsupported (NO). Answer only as a YES or NO.
c1: {c1}
c2: {c2}

E Stylistic Differences between Technical and Nontechinal Abstracts

Figure A7 shows stylistic differences between technical and nontechnical abstracts.

F Evaluation results for NSF-SciFy-20K

Tables A5, A6, and A7 summarize the results for the three generation tasks defined in Section 5 on NSF-SciFy-20K.

Model Metric	Base	Finetuned
BERTScore-F1	0.8514 ± 0.0003	0.8500 ± 0.0006
BERTScore-Precision	0.8515 ± 0.0003	0.8513 ± 0.0007
BERTScore-Recall	0.8516 ± 0.0003	0.8496 ± 0.0005
ROUGE-rouge1	0.3351 ± 0.0013	0.3141 ± 0.0023
ROUGE-rouge2	0.0705 ± 0.0008	0.0936 ± 0.0016
ROUGE-rougeL	0.1773 ± 0.0008	0.1967 ± 0.0016
ROUGE-rougeLsum	0.1982 ± 0.0010	0.1998 ± 0.0016

Table A5: Technical to Non-Technical Abstract Task: Mistral-7B

Model	Base	Finetuned
Precision	0.4146 ± 0.0025	0.7526 ± 0.0027
Recall	0.8141 ± 0.0026	0.7354 ± 0.0026
F-score	0.5247 ± 0.0025	0.7268 ± 0.0023

Table A6: Abstract to Claims Task: Mistral-7B

Model	Base	Finetuned
Precision	0.6222 ± 0.0038	0.7219 ± 0.0027
Recall	0.6364 ± 0.0034	0.7359 ± 0.0029
F1-score	0.5668 ± 0.0033	0.7039 ± 0.0026

Table A7: Abstract to Investigation Proposals Task: Mistral-7B

G Examples of Extracted Claims and Investigation Proposals

Tables A8 and A9 provide a sampling of the extracted claims and investigation proposals.

H Examples of Scientific Claim and Investigation Proposal Categories

Please see Table A10 and A11 for the examples.

I Error Analysis Examples

I.1 Claims

Of the three proposed tasks, we consider the claim extraction task as a canonical task for performing error analysis. To do so, we consider another 120 awards from the test portion of NSF-SciFy-MatSci and NSF-SciFy-20K. These were stratified samples across the five areas of interest (similar to Section 4.3). We then generate the claims using a Mistral-7B model finetuned on NSF-SciFy-20K, resulting in 802 claims. A careful examination revealed around 2.6% of the generated claims were incorrect. To dive deeper, we categorized the erroneous claims into 5 categories. We list them here with examples:

1. Overconfidence: The claim can be overconfident about information that has qualifiers in the supporting document text (award abstract).

Award ID: 9820570

Extracted Claim: The research areas include knot theory, immiscible fluids and geodesic nets, ergodic theory, commutative algebra and vector-valued forms.
Analysis: The abstract states 'probably in the areas of,' indicating potential areas, not certainty.

2. Mixing Information: The claim can mix information from two sentences together to form wrong information.

Award ID: 1205671

Extracted Claim: The SEAQUEST experiment at Fermilab has successfully measured the asymmetry of up and down anti-quarks in the nucleon.
Analysis: The abstract mentions that SEAQUEST will follow the successful E866 measurement with more precise data, and thus it does not say SEAQUEST has already successfully measured that, but the success is describing the previous E866.

3. Overgeneralization: The claim can overgeneralize what the supporting document implies.

Award ID: 0957482

Extracted Claim: The methodology is potentially environmentally benign.
Analysis: The abstract mentions non-dangerous chemicals but does not specifically state that the methodology is environmentally benign.

4. Information Omission: The claim might omit important information from the abstract and thus the meaning is changed.

Award ID: 9409461

Extracted Claim: Frequency-domain techniques can display trade-offs between output performance and sensitivity reduction.
Analysis: The claim frames output performance and sensitivity reduction as two separate quantities and leaves out bandwidth, so it does not accurately reflect the abstract.

5. Hallucinations about Administrative Metadata: The model can sometimes hallucinate claims regarding where the funding is from and which institutions are included. While hallucination is a serious issue, it is worth noting that for this dataset and model scientific claims seem to be rarely hallucinated. In our study, all hallucinations were connected with administrative metadata.

Award ID: 0542751

Claim: The award is funded under the American Recovery and Reinvestment Act of 2009 (Public Law 111-5).
Reasoning: This claim is not mentioned in the abstract.

To mitigate these errors, uncertainty calibration and prompting strategies can reduce overconfidence and overgeneralization, encouraging the model to reflect source qualifiers. Fine-tuning with more annotated data and enforcing stricter alignment between claims and source text can address mixing information and omission issues. Retrieval-augmented generation and chain-of-thought prompting may also promote better grounding. For hallucinations about administrative metadata, entity verification or output constraints based on structured data can help. Combining these approaches with human-in-the-loop evaluation might further improve claim extraction reliability.

We performed a similar error analysis on claims extracted from Claude (See section 3.3). Our findings revealed a smaller error-rate (2.1% as opposed to 2.6%), and of the only 10 erroneous claims, 5 were hallucinations of administrative data.

I.2 Investigation Proposals

We additionally performed an error analysis on investigation proposals, following the same procedure as for claims Among 120 awards from the test portion of NSF-SciFy-MatSci and NSF-SciFy-20K, and generated the investigation proposals using a Mistral-7B model finetuned on NSF-SciFy-20K, resulting in 833 proposals. A careful examination revealed around 2.4% of the generated investigation proposals were incorrect. To dive deeper, we categorized the erroneous proposals into 4 categories. We list them here with examples:

No investigation proposals. The abstract itself does not have investigation proposals and the model forcefully generates some that are not proposals.

Award ID: 1642020

Investigation Proposal: The conference aims to advance understanding in cosmology through the presentation and discussion of new research findings.
Analysis: It implies discussion and presentation rather than a forward-looking research goal, and the abstract only seeks funding to cover junior students' registration fees, so it contains no actual investigation proposals.

Content Mismatch Error. The investigation proposal does not accurately reflect the information in the abstract—either by introducing concepts not mentioned, omitting key elements, or misrepresenting the scope or focus of the abstract's content.

Experimental Technique and Tool Development: Develop new experimental tools and techniques for data collection or experimentation.

Investigation Proposal: Access and design different novel nano-motion components.
Reasoning: The abstract focuses on designing and characterizing catalytic nanomotors with different dynamic behaviors, but it does not mention accessing or broadly designing novel nano-motion components.

Overspecification. The proposal extracted is more specific than what is actually mentioned in the abstract, containing non-existing details.

Experimental Technique and Tool Development: Develop new experimental tools and techniques for data collection or experimentation.

Investigation Proposal: Design and integrate novel nanomachines.
Reasoning: The abstract focuses on designing and characterizing nanomotors, not on integrating nanomachines. "Integrated nanomachinary systems" is mentioned only as a motivation, not as a research activity.

Existing Work. The claim is about existing work instead of a forward-looking statement.

Algorithm/Method Development: Propose novel algorithms or procedures for solving specific problems or improving performance.

Investigation Proposal: Apply the developed approach to find subtle errors in non-trivial designs.
Reasoning: The proposal describes a planned research activity, whereas the abstract only refers to a past application, not a proposed investigation.

J Potential Risks

NSF-SciFy opens new opportunities for large-scale scientific text analysis, but responsible use is important. As the dataset is automatically constructed, some extraction errors or omissions may remain, underscoring the need for careful validation in downstream applications. Its coverage of NSF award abstracts may reflect domain-specific language and institutional styles, which can inform analyses but may also introduce biases if not accounted for. Finally, while the dataset enables powerful new capabilities, users should ensure appropriate use to avoid generating or disseminating unverified claims.

K AI Writing/Coding Assistance Disclosure

In accordance with the ACL Policy on AI Writing Assistance[10], the authors attest that we used generative AI tools for assistance purely with the language of the paper, including spell checking, grammar fixes, and proof reading. Additionally, we used GPT-4o to fix LaTeX issues, and to generate LaTeX tables from spreadsheets. In all such uses, the outputs were verified by the first author for correctness.

[10] https://www.aclweb.org/adminwiki/index.php/ACL_Policy_on_Publication_Ethics#Guidelines_for_Generative_Assistance_in_Authorship

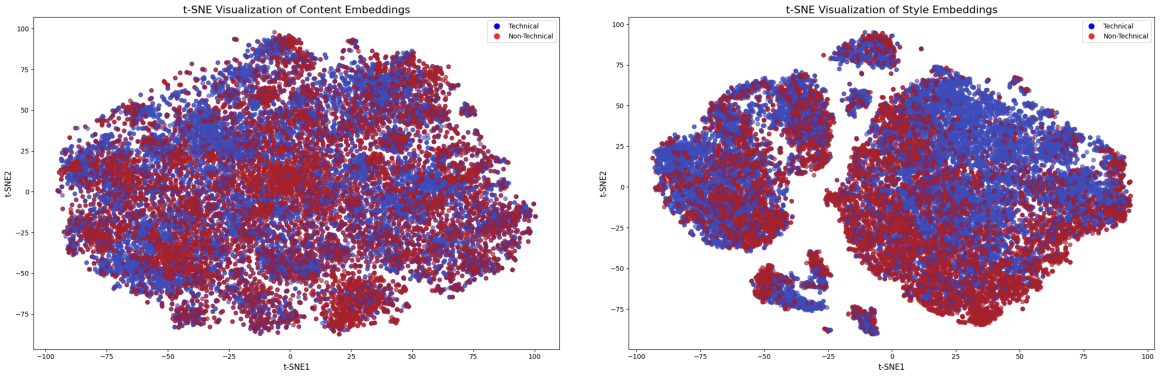

Figure A7: The t-SNE plot of comparing content embeddings from SPECTER (Cohan et al., 2020) and style embeddings from STEL (Patel et al., 2025) for technical and non-technical abstracts in NSF-SciFy-MatSci. The somewhat clear separation between technical and non-technical abstracts when using style embeddings indicate marked stylistic differences between the two kinds abstracts.

Award ID	Title	Extracted Claims
2324035	DMREF: Developing and Harnessing the Platform of Quasi-One-Dimensional Topological Materials for Novel Functionalities and Devices	["Topological insulators are electrically insulating in the bulk but host conductive surface states that are immune to impurities.", "Current TI materials face critical challenges that limit their potential.", "Quasi-1D structures promise to overcome challenges faced by current TI materials.", "Most identified topological insulators are either strongly bonded bulk materials or layered van der Waals materials.", ...]
9814055	Kinks and Surface Potentials	["Atomically flat terraced surfaces for thin TEM samples can be prepared under moderate (10-7 Torr) vacuum conditions by annealing in oxygen or vacuum for materials such as sapphire, SiC and MgO.", "\"Forbidden\" Bragg reflections arise from the stacking fault between partial dislocations.", "The surface potential is critical for chemical reactions at surfaces, adsorption, catalysis, epitaxy, diffusion bonding process, oxidation, and semiconductor crystal growth.", ...]
0821136	MRI: Acquisition of an Imaging Spherical Aberration Corrector and a Lorentz Lens for Magnetic Materials Characterization	["The attainable spatial resolution of uncorrected Lorentz instruments is in the range 10-15 nm.", "Delocalization effects cause significant image blurring in uncorrected Lorentz microscopes.", "Recent developments in aberration correction make it possible to correct the spherical aberration of a Lorentz lens.", "The size of written bits in state-of-the-art magnetic recording media is comparable to the magnetic resolution of uncorrected Lorentz microscopes.", "Transmission electron microscopes have suffered from lens aberration since their invention in the 1930s.", "The Hubble space telescope suffered from a similar aberration when first launched.", ...]

Table A8: A sample of extracted claims from the NSF-SciFy-MatSci dataset. Award IDs are hyperlinked to the NSF's Award database.

Award ID	Title	Extracted Investigation Proposals
2324035	DMREF: Developing and Harnessing the Platform of Quasi-One-Dimensional Topological Materials for Novel Functionalities and Devices	["Predict, design, synthesize, and control topological phases in quasi-1D topological materials.", "Design and demonstrate emergent materials, functionalities, and devices, including moir\'e quasi-1D TIs, stable and high temperature quantum spin Hall (QSH) insulators, and quantum intelligent sensors.", "Expand research to include other selected quasi-1D materials families through collaborations.", "Discover or realize novel topological materials and phases.", "Study topological phase transitions and control.", ...]
9814055	Kinks and Surface Potentials	["Observe dislocation kinks by atomic resolution TEM in materials such as sapphire, SiC and MgO.", "Use \"forbidden\" Bragg reflections to form lattice images without surface noise.", "Determine which process (kink formation, kink migration or obstacles along the dislocation line) limits kink (and hence dislocation) velocity, for given conditions of temperature and stress.", "Extend quantitative convergent-beam TEM measurements of bonding in crystals to the RHEED geometry to refine the electrostatic potential extending into the vacuum from ceramic surfaces.", "Measure modifications to the surface potential resulting from the deposition of a monolayer or more of atoms.", ...]
0821136	MRI: Acquisition of an Imaging Spherical Aberration Corrector and a Lorentz Lens for Magnetic Materials Characterization	["Acquire an imaging spherical aberration corrector and a Lorentz lens for magnetic materials characterization.", "Add these components to an existing FEI Titan 80-300 TEM.", "Bring the spatial resolution in Lorentz mode down to less than 1 nm, with negligible delocalization effects.", "Enable direct quantitative study of magnetic features at a length scale of around 1 nm .", "Obtain new scientific results on material systems for which these observations were previously impossible.", "Impact a large number of research groups within CMU, as well as collaborations with local industry and several national laboratories.", ...]

Table A9: A sample of extracted investigation proposals from the NSF-SCIFY-MATSCI dataset. Award IDs are hyperlinked to the NSF's Award database.

Category: Capability/Application of Technology/Method
Memory-centric computing capitalizes on extensive parallelism in memory arrays.
The Illinois group has joined the fixed target COMPASS experiment at CERN.
An electronics company is involved in the project, making imaging products in this energy regime.
Category: Definition/Classification
The RV Weatherbird II is owned and operated by the Bermuda Biological Station for Research (BBSR), Inc.
The program will include topics such as dark matter, dark energy, inflation, and gravitational waves.
The shear zone in question is the Cuyamaca-Laguna Mountains shear zone.
Category: Statement of Problem/Knowledge Gap
Current efforts on analyzing tree-informed compositional data are primarily designed for individual applications.
CU began the Guerrero GPS project in 1997.
High pressure-low temperature metamorphism is often obscured by post-tectonic thermal equilibration or later deformation and mineral growth.
Category: Experimental Result/Finding/Measurability
Lattice QCD has made important progress.
RBP repression is absent when an oncoprotein is present.
Over 100 of 650 U.S. electronics fabricators have gone out of business in the past five years, according to a 1999 White Paper by the Interconnection Technology Research Institute.
Category: Established Scientific Fact/Principle
Dynamic programming includes well-known search algorithms like breadth-first search, Dijkstra's algorithm, A*, value iteration and policy iteration for Markov decision processes.
The electron carries a magnetic moment.
Stars in clusters evolve off the main sequence, become red giants, and ultimately horizontal branch stars.
Category: Observed Phenomenon/Property
The lake level of Laguna Paron was artificially lowered in 1985.
Laminated sediments are exposed in Laguna Paron, Peru.
The study sites exhibit extreme differences (1 to 2 orders of magnitude) in larval settlement.
Category: Process/Mechanism Description
Exciton-phonon and exciton-exciton interactions contribute to decoherence at finite temperatures.
The fidelity of translation is determined by the accuracy of aminoacyl-tRNA selection by ribosomes and synthesis of cognate amino acid/tRNA pairs by aminoacyl-tRNA synthetases.
The evaluation process includes both direct and indirect measures of student success and learning.
Category: Hypothesis/Theoretical Prediction
Assemblages that combine human-technology partnerships are stronger than individual humans or machines.
Mating advantage in guppies appears to result from female sexual responses to unusual males.
The long wavelength part of the CBR spectrum is important for constraining the evolution of the intergalactic medium.

Table A10: Scientific claim categories found in NSF-SCIFY and 3 randomly selected examples for each category.

Category: Academic Training and Curriculum Development
Develop a generic geometric interpretation to the wavelet frame transform by studying its relations with differential operators within various variational frameworks.
Support participation in the visitor program activities during 2018 - 2020.
Measure the contributions of antiquarks to nucleon spin using the PHENIX polarized pp program with an Illinois-led muon trigger upgrade.
Category: Experimental Technique and Tool Development
Develop a generic geometric interpretation to the wavelet frame transform by studying its relations with differential operators within various variational frameworks.
Measure the contributions of antiquarks to nucleon spin using the PHENIX polarized pp program with an Illinois-led muon trigger upgrade.
Develop a method of creating sulfur ylides with improved yields.
Category: Theoretical Analysis and Computational Modeling
Develop a generic geometric interpretation to the wavelet frame transform by studying its relations with differential operators within various variational frameworks.
Deepen understanding about how to recognize the complexity of certain types of computational problems.
Support participation in the visitor program activities during 2018 - 2020.
Category: Human/User Study
Focus on the settling and juvenile stages of 7 dominant species within subtidal marine epifaunal communities along the coast of southern New England.
Examine the impact of sea ice on the distribution and abundance of zooplankton.
Examine and model visual tracking of continuously moving targets in normal human subjects.
Category: Algorithm/Method Development
Develop a generic geometric interpretation to the wavelet frame transform by studying its relations with differential operators within various variational frameworks.
Deepen understanding about how to recognize the complexity of certain types of computational problems.
Develop a method of creating sulfur ylides with improved yields.
Category: Policy/Guidelines/Standards Work
Design, fabricate, assemble, align, test, integrate, and calibrate a sensitive CCD camera system.
Provide funding to offset registration fees for about 12 graduate students or postdocs at the COSMO-16 conference.
Replace two semi-conductor detectors in the Neutron Activation Laboratory.
Category: Interpretability/Alignment Analysis
Understand and correct for hidden assumptions in Bayesian inference algorithms.
Develop assemblages for human-technology partnerships in visually based cognition-oriented tasks in radiology.
Systematically investigate and proactively prevent specious configurations.
Category: Deployment/Field Study
Develop a method of creating sulfur ylides with improved yields.
Design, fabricate, assemble, align, test, integrate, and calibrate a sensitive CCD camera system.
Measure chlorofluorocarbons (CFC-11, CFC-12, CFC-113) on the 26 degrees N transect in winter 2004.
Category: Tooling/Systems/Infrastructure
Deepen understanding about how to recognize the complexity of certain types of computational problems.
Design, fabricate, assemble, align, test, integrate, and calibrate a sensitive CCD camera system.
Understand and correct for hidden assumptions in Bayesian inference algorithms.
Category: Empirical Benchmarking/Evaluation
Measure the contributions of antiquarks to nucleon spin using the PHENIX polarized pp program with an Illinois-led muon trigger upgrade.
Obtain accurate colors and brightnesses of the brighter stars in 50 globular clusters over a two-year period.
Develop a new density cumulant functional theory.

Table A11: Investigation proposal categories found in NSF-SCIFY and 3 examples for each category.