Zero-shot Scientific Claim Verification Using LLMs and Citation Text

Carlos Alvarez* **Maxwell Bennett*** **Lucy Lu Wang**
University of Washington
{calvar, maxmiles, lucylw}@uw.edu

Abstract

Due to rapidly changing and advancing science, it is important to check the veracity of scientific claims and whether they are supported by research evidence. Previous versions of this task depended on supervised training, where labeled datasets were constructed through manual claim writing and evidence identification, sometimes coupled with mining citation relationships in papers. In this work, we investigate whether zero-shot scientific claim verification could be enabled using large language models (LLMs) and distant supervision examples taken directly from citation texts. We derive an in-context learning (ICL) dataset, SCitance, consisting of citation sentences ("citances"), LLM-generated negations, evidence documents, and veracity labels, and find that prompting GPT-4 with ICL examples from this dataset yields comparable performance (within 1 point F1) to previous finetuned models trained on manually curated claim-evidence pairs. Our results suggest that prompting LLMs with citance-evidence pairs directly poses a viable alternative to finetuning scientific claim verification models with manually-curated data.[1]

1 Introduction

Verifying scientific claims is important for assessing the rigor of the research enterprise and for addressing concerns such as misinformation or misinterpretation of scientific output. Prior work in scientific claim verification relies on manually annotated supervision data (expert-written claims verified against documents) (Wadden et al., 2020; Sarrouti et al., 2021; Saakyan et al., 2021), which can be expensive to curate and difficult to scale. This has inspired work in zero- or few-shot settings, which have demonstrated success by extracting claims using supervised models, then training claim verification models on the extracted claims (Pan et al., 2021; Wright et al., 2022). Here, we investigate two questions: (i) whether expert-written claims are needed at all, and (ii) whether large language models (LLMs) can verify scientific claims in zero- or few-shot settings with no need for supervision labels or model finetuning.

Towards (i), we investigate whether citation sentences ("citances") and citing relationships from the scientific literature could be used directly as noisy labeled data, without converting citances to claims. We construct a dataset of citance-evidence pairs, SCitance, based on the claim-evidence pairs from SciFact (Wadden et al., 2020), and use these as in-context learning (ICL) examples for prompting. To derive contradictory examples, we use GPT-3.5 to generate negations of citances. Towards (ii), we design zero- and few-shot prompts for claim verification using GPT-3.5 and GPT-4, and find that GPT-4 with ICL performs comparably at abstract-level verification compared to supervised models trained on expert-annotated claim-evidence pairs (within 1 point F1). This result indicates that contemporary LLMs could be adapted to perform scientific claim verification with few supervision labels, which could dramatically lower the costs associated with domain transfer of these models.

2 Related Work

Scientific Claim Verification Since the introduction of large-scale fact verification datasets such as FEVER (Thorne et al., 2018) and UKP Snopes (Hanselowski et al., 2019), notable datasets supporting claim verification in the scientific domain have followed (Wadden et al., 2020; Saakyan et al., 2021; Sarrouti et al., 2021; Kotonya and Toni, 2020). We base our work on SciFact (Wadden et al., 2020), a dataset of 1,400 expert-written biomedical claims paired with evidence abstracts, which using citances as a source of claims and their corresponding evidence relationships. In this work, our focus is exploring LLMs for verifying scien-

*denotes equal contribution
[1]Code and data at https://github.com/larchlab/scitance

tific claims using citation contexts from existing literature directly.

Since the emergence of LLMs, many works have explored prompting for fact/claim verification. Zhang and Gao (2023) experimented with ICL for news claim verification, finding that prompting with a few examples achieves performance comparable to that of supervised models. They and others (Wang and Shu, 2023) also demonstrated the ability of LLMs to support explainable claim verification, where rationales are provided alongside veracity judgements or grounded to knowledgebases. LLMs have also been used to coordinate fact-checking of complex claims (Pan et al., 2023), which are similar to citances in scientific text.

Negation generation Prior work has investigated how to automatically generate negations to train claim verification systems. Pan et al. (2021) introduce QACG, which automatically generates QA pairs from Wikipedia and converts these into supporting, contradicting, or unrelated claims. Wright et al. (2022) use an knowledge-based entity substitution approach, KBIN, which substitutes biomedical entities with related entities from the UMLS knowledgebase. In their analysis, they find that KBIN most often produces claim variations, rather than true negations.

Alternatively, Saakyan et al. (2021) introduce a novel approach to negation generation that uses masked language model infilling. After generating several negations of salient words, the method selects negations with the highest contradiction score using a RoBERTa model (Liu et al., 2019) trained on Multi-NLI (Williams et al., 2018). However, this method requires some human supervision to optimize contradiction score thresholds, and can only modify a claim by changing a single word or multi-word expression. In our work, we investigate whether LLMs can generate negations; LLMs do not rely on specific knowledgebases and can produce multiple modifications to the same sentence, which is necessary for negating citances with complex syntax and multiple clauses.

3 Dataset

We derive our dataset, SCitance, from SciFact (Wadden et al., 2020), a dataset of 1.4K manually-written scientific claims from citances in biomedical papers verified against over 5000 evidence documents. For each claim-evidence pair, trained annotators provided labels for whether the evidence supports, refutes, or offers insufficient information (NEI) towards the claim. To validate whether citance-evidence relationships can be used to train scientific fact checking models directly, without the need to extract and rewrite claims, we make use of the original citances in SciFact (mapping provided by the authors) rather than the rewritten claims to train our models.

Fold	Support	Refute	NEI	All
Train	178	155	134	467
Dev	38	31	29	98
Test	35	39	17	91
All	251	225	180	656

Table 1: Distribution of labels in SCitance

We keep all citance-evidence pairs from the train and dev set with SUPPORT and NEI labels, yielding 251 supporting and 180 NEI entries. These numbers are notably lower than SciFact's due to: (i) in SciFact, multiple claims could be written based on a single citance; and (ii) each claim from the same citance with the same supporting evidence document translates to only one instance in our dataset. This also carries over to NEI examples.

Generating citance negations Mapping claims to citances only yields SUPPORT and NEI-labeled training instances, yet contradictory claim-evidence relationships are necessary to train a claim verification model. While claim negations were written manually for SciFact to produce contradictory training samples, subsequent analysis has shown that this process may introduce lexical biases into negations that could be used by models to shortcut predictions (Wadden et al., 2020; Wright et al., 2022). To offset the cost of producing manual negations and mitigate lexical biases, we use GPT-3.5[2] to automatically negate citances. Negating citances remains a non-trivial task. Citances from scientific papers have complex syntactic structure, offering multiple correct ways to negate the content, where some negations need to be reflected as multiple changes across the entire sentence.

To identify effective prompt instructions for generating negations, we examine prompt variants for both GPT-3.5 and GPT-4 and compare results on 20 citances sampled from SCitance. We evaluate the resulting negations on several criteria: (i) the negation should maintain all clauses in the origi-

[2]The model we used to generate negations in SCitance, `text-davinci-003`, has since been deprecated.

Includes NEI	Input	Model setting	Micro-F1	Macro-F1
Yes	Citance-abstract pairs	MultiVerS*	73.8	73.6
Yes	Citance-abstract pairs	Zero-shot GPT-3.5	44.2	34.5
Yes	Citance-abstract pairs	Few-shot GPT-3.5	43.7	36.5
Yes	Citance-abstract pairs	Zero-shot GPT-4	80.1	79.1
Yes	Citance-abstract pairs	Few-shot GPT-4	75.4	73.3
No	Citance-abstract pairs	Zero-shot GPT-4	88.1	88.1
No	Citance-abstract pairs	Few-shot GPT-4	86.8	86.7
No	Citance-only	Zero-shot GPT-4	69.7	68.4
No	Citance-only	Few-shot GPT-4	71.6	70.9

Table 2: SCitance test set performance. All experiments are conducted in the abstract-provided setting (where the gold abstract is provided as input), except for citance-only (in which no evidence abstracts are provided). *MultiVerS was trained on claim-abstract pairs.

nal citance, (ii) the negation should be active, not passive (i.e., simply inserting a negation word like "not"), and (iii) only the main claims made in the citance should be negated. Our evaluation found that the instruction "Please negate this sentence by changing as few words as possible in the original sentence" yielded the best results.

Upon comparing model outputs, we find that GPT-4 tends to negate by inserting negation words like "not" rather than flipping the meaning of words in the sentence. For example, with the citance "Approximately 90% of SIDS deaths occur in infants aged less than 6 months old," GPT-4 changed "...[do not] occur..." whereas GPT-3.5 changed "90%" to "10%." Based on these findings, we elect to use GPT-3.5 to generate negations. We implement two additional checks to improve the quality of negations. First, we only keep negations within $\pm$10% of the original citance token length.[3] Second, we verify successful negation by feeding the original and negated citance into GPT-4 and asking the model to determine whether they are proper negations of one another.

Full prompts and additional evaluation examples are provided in Appendix B. The final 251 SUPPORT, 225 REFUTE, and 180 NEI instances in SCitance are split into train, dev, and test sets. Label distributions are provided in Table 1.

4 Experiments

We adopt the abstract-level scientific claim verification task definition from Wadden et al. (2020). Given a claim and retrieved evidence abstract, the goal is to determine whether the evidence supports, refutes, or offers insufficient information towards the claim. Here, instead of only claims, we provide either citances or claims as our verification objection. We report performance on SCitance and SciFact (via the public leaderboard[4]).

Models We compare prompting methods against pretrained language models on this task. MultiVerS (Wadden et al., 2022) is a supervised model that uses the Longformer encoder (Beltagy et al., 2020) to create a shared encoding for claims and abstracts, with multiple classification heads to simultaneously predict the claim veracity label and extract evidence sentences. We report results from the version of MultiVerS trained on FEVER and weak supervision datasets, then fine-tuned on SciFact (Wadden et al., 2020).

All prompting experiments are conducted with two models using the OpenAI API: GPT-3.5 (`gpt-3.5-turbo-0125`) and GPT-4 (`gpt-4-0613`) (OpenAI, 2023). Temperature was set to 0.2 and no limit was placed on maximum tokens.

Prompt settings We prompt models in zero- and few-shot ICL settings (prompts in Appendix A). For few-shot experiments, we use similar instructions as in the zero-shot setting, but include examples from the train split of SCitance or SciFact depending on the experiment. We randomly select an example corresponding to each label (SUPPORT, REFUTE, NEI) to include before providing the test sample. We report model performance on abstract-level classification for SCitance and SciFact.

Retrieval setting When evaluating on SCitance, we report performance in the abstract-provided setting, without incorporating a retrieval

[3] In rare cases, entire clauses are removed in the negation.

[4] https://leaderboard.allenai.org/scifact/

step. When evaluating on SciFact via the leaderboard, we employ dense retrieval using Sentence-BERT (Reimers and Gurevych, 2019) with the `S-PubMedBert-MS-MARCO-SCIFACT` checkpoint as implemented by Deka et al. (2022). We retrieve the top 3 documents per claim and use these as evidence abstracts in prompting.

5 Results

Performance on SCitance Model performance on SCitance is provided in Table 2. Surprisingly, zero-shot prompting with GPT-4 yields the best results on SCitance (80.1 micro-F1), markedly higher than MultiVerS trained on claim-abstract pairs and several points better than the few-shot setting with in-context examples (75.4 micro-F1). This indicates that GPT-4 is able to reason about citation-evidence relationships out-of-the-box and without modifying citances into atomic claims. By contrast, we observe significantly worse performance from zero- and few-shot GPT-3.5, with micro-F1 scores of 44.2. and 43.7. For MultiVerS, this task variant represents a domain shift, and the model, having been trained on claim-abstract pairs, performs less well when given citances.

Performance on SciFact On the SciFact test set, we see comparable performance between few-shot prompting with GPT-4 and MultiVerS (Table 3). Using SciFact's own training data as in-context learning examples provided a marginal boost to performance over using citances. In contrast to performance on SCitance, zero-shot prompting with GPT-4 performed less well—by a difference of 3 points. GPT-3.5, however, performs far worse in both zero- and few-shot settings, similar to its performance on SCitance. Nonetheless, these results suggest that citance-abstract relationships may provide comparable in-context supervision to claim-abstract relationships (as shown in GPT-4 results), and claim rewriting may be unnecessary when generalizing these methods to other domains.

Ablations We conduct citance-only experiments to assess biases introduced by our negation generation procedure. We retain only citances with support or refute labels and prompt GPT-4 for the veracity of each citance. Table 2 shows zero- and few-shot results comparing using only citances as input against using citance-abstract pairs. Low F1-scores associated with citance-only settings indicate that negation generation did not introduce significant biases into the data that can be exploited by LLMs for claim verification.

Model	Setting	F1
MultiVerS	Trained on claims	72.5
GPT-3.5	Zero-shot	35.0
	Few-shot (ICL w/ citances)	39.0
	Few-shot (ICL w/ claims)	39.9
GPT-4	Zero-shot	68.6
	Few-shot (ICL w/ citances)	71.7
	Few-shot (ICL w/ claims)	72.2

Table 3: Performance on SciFact test set

6 Discussion & Conclusion

This study explores zero-shot scientific claim verification using LLMs and citation texts directly, demonstrating effective transfer to verifying scientific claims. GPT-4, when prompted with ICL examples from SCitance, performs within 1-pt F1 of finetuned models that rely on manually curated claim-evidence pairs. Thus, using citation sentences directly as noisy labeled data and prompting LLMs to produce small numbers of negations and counter-examples can serve as a potential alternative to manual data curation. Such datasets would be much easier to create in a novel scientific domain due to the common occurrence of citation relationships in scientific literature that could be used to bootstrap annotations.

However, there are limitations to our approach. Negating or generating variants of complex sentences while preserving logical internal relationships remains a challenge. Our work also does not contend with explainability, an important facet of claim verification that has received ample attention in recent years. While we did not conduct experiments to this effect, prior efforts in news and general domain claim verification (Zhang and Gao, 2023; Wang and Shu, 2023) suggest that LLMs excel at rationalizing verification decisions, which could also be tested in the scientific domain.

SCitance was based on claim-evidence relationships validated by annotators for SciFact, so they are not truly devoid of any manual curation. Follow-up work should investigate whether a dataset like SCitance could be constructed directly from citation relationships from independent sets of scientific documents. This would demonstrate the feasibility of citation-based dataset construction to support task transfer to other scientific disciplines.

Limitations

As discussed, our approach faces several limitations. Prompt engineering was crucial for performance, yet minor changes in prompt instruction can lead to signficant changes in performance. Even with our best-performing prompts, the model would still fail in some cases to contend with the complex structure of naturally occurring citances. Our study also does not address the important explainability aspect of claim verification. Finally, the demonstrated effectiveness of LLMs is limited to one dataset in a single scientific subdomain, and the dataset used was not entirely free from manual curation since it is based on SciFact. The involvement of other datasets and domains, as well as constructing other such datasets automatically from scratch would be opportunities for future work.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Pritam Deka, Anna Jurek-Loughrey, and P Deepak. 2022. Improved methods to aid unsupervised evidence-based fact checking for online health news. *Journal of Data Intelligence*, 3(4):474–504.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

OpenAI. 2023. GPT-4 Technical Report.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot fact verification by claim generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 476–483, Online. Association for Computational Linguistics.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. 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.

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.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. Multivers: Improving scientific claim verification with weak supervision and full-document context.

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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2448–2460, Dublin, Ireland. Association for Computational Linguistics.

Xuan Zhang and Wei Gao. 2023. Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1011, Nusa Dua, Bali. Association for Computational Linguistics.

A Prompts

Prompts used for all experiments are reproduced in full in Table 4. Experiments on SciFact use the same prompts as the associated settings for SCitance, with test claims instead of citances.

B Negation Prompts

We test six prompts on twenty citances using both GPT-3.5 and GPT-4, and compare the resulting negations to determine the optimal setting for negation generation. Through qualitative evaluation, we find that "Please negate this sentence by changing as few words as possible in the original sentence" yields the best results.

We evaluate negations on the following criteria: (i) the negation should maintain all clauses in the citance, (ii) the negation should be active, not passive (e.g., inserting a negation word like "not"), and (iii) only the main claim made in the citance should be negated. For (iii) for example, with a more complex citance such as "Such sequence variation is likely to consist of rare variants, present in less than 1% of the population, with potentially larger penetrance effects than previously identified common variants", it is essential that the inner dependent clause is not changed to "...present in more than..." because that clause provides context but is not the main claim in the citance.

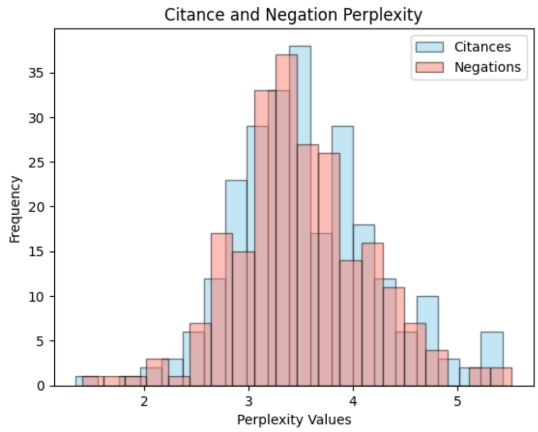

Figure 1: Distribution of GPT-2 perplexity scores associated with citances and automatically generated negations. The distributions overlap.

C Assessing lexical bias in negations

We assess lexical bias among generate negations. When predicting the verification label, certain trigger words (e.g., "not") may correlate with the contradiction label, and may be exploited by the classification model as a shortcut.

To determine if negations contain different lexical distributions to the original citances, we calculate and compare perplexity score distributions generated using GPT-2 for citances and negations (Radford et al., 2019). We conduct an independent two-tailed T-test and find that any difference between the two perplexity distributions (Figure 1) is not statistically significant ($t = 0.727$; $p = 0.468$).

NEI	Task	Setting	Prompt
No	Citance-only	Zero-shot	"Given a claim, please determine whether the existing academic literature SUPPORTS or CONTRADICTS the claim (even if you cannot reference specific abstracts). Please return your answer as only the capitalized token, as well as an explanation or rationale for the answer. Claim: {}"
No	Citance-only	Few-shot	"The following are examples of claims from a research paper and the corresponding abstract from the paper they are citing. This is an example of an abstract that SUPPORTS the claim: Supporting abstract: {} Claim: {} This is an example of an abstract that CONTRADICTS the claim: Contradicting abstract: {} Claim: {} Please obey the following: With no specific abstracts, please make an estimation whether the existing academic literature (and not the abstracts above) SUPPORTS or CONTRADICTS the claim. You must choose SUPPORTS or CONTRADICTS. Please return your answer as only the capitalized token, as well as an explanation or rationale for the answer. Claim: {}"
No	Citance-abstract pairs	Zero-shot	"Please obey the following: With a specific abstract, please make an estimation whether the abstract SUPPORTS or CONTRADICTS the claim. You must choose SUPPORTS or CONTRADICTS. Please return your answer as only the capitalized token, as well as an explanation or rationale for the answer. Abstract: {} Claim: {}"
No	Citance-abstract pairs	Few-shot	"The following are examples of claims from a research paper and the corresponding abstract from the paper they are citing. This is an example of an abstract that SUPPORTS the claim: Supporting abstract: {} Claim: {} This is an example of an abstract that CONTRADICTS the claim: Contradicting abstract: {} Claim: {} Please obey the following: given a new abstract and claim pair, please make an estimation whether the abstract SUPPORTS or CONTRADICTS the claim. You must choose SUPPORTS or CONTRADICTS. Please return your answer as the capitalized token, as well as an explanation or rationale for the answer. New abstract: {} Claim: {}"
Yes	Citance-abstract pairs	Zero-shot	"Please obey the following: With a specific abstract, please make an estimation whether the abstract SUPPORTS, CONTRADICTS, or if there is NOT_ENOUGH_INFO to determine. You must choose SUPPORTS or CONTRADICTS or NOT_ENOUGH_INFO. Please return your answer as only the capitalized token, as well as an explanation or rationale for the answer. Abstract: {} Claim: {}"
Yes	Citance-abstract pairs	Few-shot	"The following are examples of claims from a research paper and the corresponding abstract from the paper they are citing. This is an example of an abstract that SUPPORTS the claim: Supporting abstract: {} Claim: {} This is an example of an abstract that CONTRADICTS the claim:Contradicting abstract: {} Claim: {} This is an example of an abstract with NOT_ENOUGH_INFO about the claim: Missing info abstract: {} Claim: {} Please obey the following: given a new abstract and claim pair, please make an estimation whether the abstract SUPPORTS, CONTRADICTS, or if there is NOT_ENOUGH_INFO to determine. You must choose SUPPORTS or CONTRADICTS or NOT_ENOUGH_INFO. Please return your answer as the capitalized token, as well as an explanation or rationale for the answer. New abstract: {} Claim: {}"

Table 4: Model Prompts

Negation prompt
"Please provide two different examples of a negated version of the following sentence, by changing as few words as possible in the original sentence: "
"A negated sentence is a sentence that has had one or more words added, removed, or changed so that the resulting sentence has the opposite meaning from the original. Here are two examples: Original sentence: Biodegradable and biocompatible 0DBMs seem to be promising candidates to solve the problem, since they show great abilities to deliver the biomolecules in to cells, and some 0DBMs even show inductive properties themselves. Negated sentence: Biodegradable and biocompatible 0DBMs do not seem to be promising candidates to solve the problem, since they show limited abilities to deliver the biomolecules in to cells, and some 0DBMs even lack inductive properties themselves. Original sentence: Approximately 90% of SIDS deaths occur in infants aged less than 6 months. Negated sentence: Approximately 10% of SIDS deaths occur in infants aged less than 6 months. Please provide a negated version of the following sentence: "
"Please provide a new sentence with the opposite meaning as the following, by changing as few words as possible in the original: "
"Please provide a new sentence with the opposite meaning as the following, by changing a small number of words: "
"Please provide a new sentence with the opposite meaning as the following: "
"Please negate this sentence by changing as few words as possible in the original sentence: "

Table 5: Prompt variants tested for negation generation