Towards Efficient Self-Explainable Climate-Related Claim Verification with Generative Models

Siting Liang, Omar Adjali, Daniel Sonntag
German Research Center for Artificial Intelligence, Oldenburg University
Germany
{siting.liang, omar.adjali, daniel.sonntag}@dfki.de
{siting.liang, daniel.sonntag}@uni-oldenburg.de

Abstract

In this work, we present an empirical investigation into two self-explanatory inference paradigms using pre-trained language models with different sizes, based on our participation in the **ClimateCheck@NSLP 2026** shared task on climate-related claim verification. This task aims to address the increasing amount of climate misinformation and disinformation on social media, emphasizing the importance of basing claims on reliable scientific evidence. Our study investigates the impact of different explanation strategies on entailment-based verification performance in scientific claim verification, while analyzing the trade-off between reasoning complexity and computational efficiency.

Keywords: Scientific Claim Verification, Large Language Model, Natural Language Inference

1. Introduction

Scientific misinformation through digital media has been posing significant societal risks and motivating the development of automated claim verification systems (Sarrouti et al., 2021; Wadden et al., 2020; Ahmad et al., 2025). The task of scientific claim verification involves assessing whether a scientific claim is supported, refuted, or lacks sufficient evidence when compared against scientific literature. In this work, we examine two self-explanatory inference approaches, exploring how explanation strategies impact performance and the trade-off between reasoning complexity and computational efficiency in the context of verifying climate-related claims from the **ClimateCheck@NSLP 2026 shared task** (Abu Ahmad et al., 2026).

Traditional approaches to automated claim verification have primarily focused on achieving high predictive accuracy, often treating the problem as a black-box natural language inference (NLI) task (Bowman et al., 2015). Subsequent work, such as e-SNLI (Camburu et al., 2018) extended the standard natural language inference (NLI) framework by incorporating human-annotated explanations alongside entailment labels, enabling the joint learning of classification and explanation generation. This line of research advances self-explainable inference by encouraging models not only to predict entailment relations but also to provide explicit justifications for their decisions. Building upon this direction, more studies have increasingly focused on generative models to develop further self-explainable inference methods to meet the growing demand not only for accurate predictions but also for transparent and interpretable reasoning that domain experts can validate and trust (Jullien et al., 2023; Liang and Sonntag, 2025b).

Another challenge in scientific claim verification lies in long-context encoding, as the length of abstracts can easily span thousands of tokens. This exceeds the context windows of many pre-trained language models and imposes constraints on the choice of applicable models. Recent advances in large language models, many of which are capable of encoding thousands to tens of thousands of tokens, present new opportunities for developing systems that can both verify claims and generate human-readable explanations for their decisions (Wei et al., 2022). However, challenges such as hallucination and sensitivity to intermediate reasoning strategies remain significant obstacles, particularly in high-stakes scientific verification settings where faithful and evidence-grounded reasoning is essential (Xia et al., 2024; Saxena et al., 2024).

Previous studies have demonstrated that structured prompting frameworks can improve the reasoning abilities in large language LLMs (Yu et al., 2023). For instance, Liang and Sonntag (2025a) demonstrated that carefully designed multi-step reasoning prompts can improve the performance of LLMs on complex claim verification tasks in biomedical domains. However, such methods are often computationally expensive and inefficient. Despite the increased computation and processing time, the observed performance gains are minimal, suggesting that such methods may not be a practical solution for real-world verification systems. In this work, we attempt to simplify the reasoning pipeline to improve efficiency for the climate-related claim verification task. We conduct experiments to compare two explanatory paradigms, aiming to develop a more practical yet interpretable verification framework.

Proceedings of the 3rd Int. Workshop on Natural Scientific Language Processing (NSLP 2026) @ LREC 2026, pages 255–260
12 May 2026.

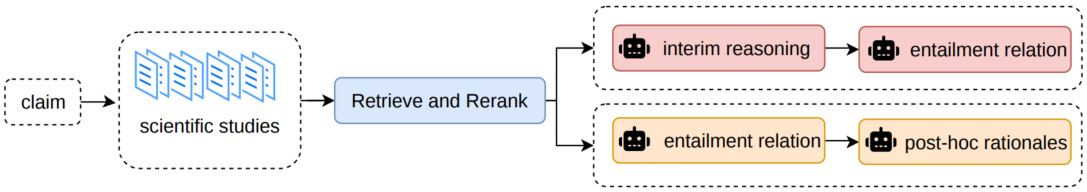

Figure 1: Overview of two-stage claim verification workflow with two explainability paradigms.

2. Approach

The climate-related claim verification workflow consists of two main stages: first, retrieving contextually relevant scientific abstracts from the corpus, and second, performing entailment-based verification between each claim-abstract pair. In the first stage, we follow the baseline procedure outlined in (Ahmad et al., 2025), using BM25[1] for initial retrieval, followed by a cross-encoder[2] to re-rank the results and identify contextually relevant scientific articles for each claim. For each claim–abstract pair, a pre-trained generative large language model is used to infer the entailment relationship ('Supports', 'Refutes', or 'Not Enough Information') using two alternative paradigms that are self-explanatory, as illustrated in Figure 1.

- **Intermediate Reasoning (IR),** where explanations are generated **prior to predicting** the entailment label, reflecting the model's step-by-step reasoning process.
- **Post-hoc Rationalization (PR),** where explanations are produced **after** the entailment label has been predicted, justifying the model's decision.

Our experiments are designed to address the following research questions:

1. How do different explainability paradigms impact the performance of large language models on complex reasoning tasks in scientific claim verification?
2. How does model size influence the effectiveness of each explainability paradigm in this context?

In our experiments, we evaluate a combination of lightweight open-source models and a low-cost commercial model to balance reasoning capability and efficient reproducibility, listed in Table 1.

GPT-4o-mini is a cost-efficient variant of the GPT-4 series that retains strong reasoning performance while being more practical for large-scale experimentation. In addition, we incorporate instruction-tuned open-source (Mistral-Nemo and Phi-3-Medium), which have demonstrated competitive performance on previous related work (Liang and Sonntag, 2025a). These mid-sized models provide a suitable balance between computational efficiency and reasoning capacity. We design separate prompting templates for the intermediate reasoning and post-hoc rationalization paradigms, as shown in Table 2.

In addition, we experiment with using GPT-4o-mini to generate augmented post-hoc rationalization samples based on the annotated entailment labels in the training set. These generated explanations are then used to fine-tune a much smaller generative model, such as T5-base[3] (Raffel et al., 2020), which is evaluated under both paradigms—prediction-first and prediction-after-explanation. This lightweight model requires less than 20 minutes to process the 1760 claim–abstract pairs at inference time, substantially reducing computational cost compared to large language models, which require over 200 minutes under the same setting.

Type	Parameters	Context Window
GPT-4o-mini (Achiam et al., 2023)	–	128K
Phi-3-Medium (Abdin et al., 2024)	∼14B	4K
Mistral-Nemo (Jiang et al., 2023)	12B	8K

Table 1: Comparison of the LLMs used in our experiments. We include both proprietary and open-source models to analyze the impact of model architecture and size on entailment-based claim verification under different explanatory paradigms.

Paradigm	Instruction Prompt
IR	*Given the Claim: '{claim}'* *and the abstract: '{abstract}'* *Summarize in short, is there enough information to determine if the claim aligns with the data points or if the data refutes the claim?* *→ Predict the 'logical relation' between the claim and the provided data. Select one label from ['not enough information', 'supports', 'refutes']. Return only valid JSON.*
PR	*Given the Claim: '{claim}'* *and the abstract: '{abstract}'* *Predict the 'logical relation' between the claim and the provided data. Select one label from ['not enough information', 'supports', 'refutes'].* *→ Explain your decision with evidence. Summarize in short.*

Table 2: Instruction prompts for the two explanatory paradigms with LLMs used in entailment-based claim verification.

[1] https://pypi.org/project/rank-bm25/

[2] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2

[3] https://huggingface.co/google-t5/t5-base

3. Results

In this section, we present and discuss the results obtained through submissions to the official leaderboard of the shared task. The evaluation scores are organized according to the two subtasks. Task 1.1 (Abstract Retrieval) is assessed using Recall@K (for K = 2 and 5) and B-Pref, measuring the effectiveness of retrieving relevant scientific abstracts for each claim. Task 1.2 (Claim Verification) is evaluated using weighted F1 scores, which reflect performance across all entailment classes (Supported, Refuted, Not Enough Information).

Metric	Recall@2	Recall@5	B-pref	Overall Score
Score	0.1931	0.3524	0.4010	0.3767

Table 3: Retrieval performance metrics using BM25 with sentence transformer-based reranking on Task1-1. Recall@K measures the proportion of relevant documents in the top-K results. B-pref (Binary Preference) evaluates ranking quality based on pairwise comparisons. Overall Score represents the average performance across all metrics.

These scores presented in Table 3 suggest that the current retrieval pipeline (BM25 + cross-encoder reranking) is not retrieving enough relevant scientific abstracts. Even in the top 5, only about 35% of relevant abstracts are retrieved. The B-Pref of about 0.4 also indicates the approach often ranks non-relevant abstracts above relevant ones. These low retrieval scores are likely to negatively impact downstream entailment-based claim verification, as they lack sufficient evidence to make accurate predictions.

In this work, however, we primarily focus on investigating the effectiveness of the explanatory paradigms in Task 1.2 (claim verification).

	Intermediate Reasoning				**Post-hoc Rationalization**			
Model	Prec.	Rec.	F1	Score	Prec.	Rec.	F1	Score
GPT-4o-mini	0.68	0.65	0.63	0.98	0.71	0.65	0.62	0.97
Phi-3-medium	0.67	0.64	0.62	0.97	0.63	0.50	0.41	0.76
Mistral-Nemo	0.67	0.65	0.64	0.98	0.63	0.50	0.41	0.76

Table 4: Performance scores of task 1.2 claim verification under intermediate reasoning and post-hoc rationalization paradigms in a zero-shot setting.

We compare the performance of three models under the intermediate reasoning and post-hoc rationalization paradigms on entailment-based claim verification (Task 1.2) in Table 4. Overall, the intermediate reasoning paradigm demonstrates more stable and consistent better performance across models in the zero-shot setting. GPT-4o-mini achieves the best results, indicating that larger, more capable models can better handle reasoning-generation tasks with long text input. Phi-3-medium and Mistral-Nemo have close model sizes and perform comparably to each other. However, under the post-hoc rationalization paradigm, both smaller models experience a sharp decline in recall and F1 scores. This contrast highlights the relative effectiveness of intermediate reasoning, suggesting that generating explanations prior to prediction can better support structured and evidence-grounded decision making. Although intermediate reasoning may introduce potential risks such as hallucination due to extended generation steps, the observed improvements in verification performance indicate that the trade-off between enhanced reasoning capability and generation risk is worthwhile in this context.

Paradigm	**Precision**	**Recall**	**F1**	**Score**
IR	0.49	0.49	0.49	0.84
PR	0.59	0.56	0.54	0.89

Table 5: Task 1.2 performance of T5-base trained with GPT-4o-mini augmented explanations under two explanatory paradigms.

T5-base requires significantly less computation than larger LLMs, making it an efficient and practical choice for large-scale claim verification. However, small pre-trained language models like T5-base typically cannot generate meaningful explanations independently, in contrast to LLMs. To address this, we use GPT-4o-mini to augment the training set with post-hoc rationalization examples conditioned on annotated entailment labels. Table 5 shows the performance of T5-base trained with the GPT-4o-mini–augmented post-hoc rationalizations under the two explanatory paradigms. While intermediate reasoning tends to be more stable for LLMs in a zero-shot setting, post-hoc rationalization outperforms intermediate reasoning for T5-base across all metrics. Fine-tuning on augmented post-hoc rationalizations appears particularly beneficial for smaller generative models, improving prediction quality when explanations are produced after the model makes its initial entailment decision.

Examples of generated explanations across models and explanatory paradigms are presented in Table 6. Overall, the scores indicate that model size remains the primary factor influencing reasoning capability in LLMs. Larger models consistently perform better across both paradigms, demonstrating greater robustness in entailment-based verification and the ability to generate coherent explanations. Nonetheless, the trade-off between performance, computational efficiency, and reproducibility remains an important consideration.

Model (Type)	Intermediate Reasoning	Post-hoc Rationalization
GPT-4o-mini (Zero-shot)	`The claim distinguishes short-term weather events from long-term climate trends. The abstract explains that the polar vortex drives winter variability and interacts with patterns like the North Atlantic Oscillation and El Niño-Southern Oscillation. Although the vortex can cause cold conditions, this does not contradict long-term climate change. | Relation: supports`	`supports. The abstract describes the polar vortex as a driver of winter variability, aligning with the claim that cold weather does not negate climate change. Short-term anomalies should not overshadow long-term climate trends.`
Phi-3 Medium (Zero-shot)	`The abstract presents the polar vortex as a natural weather pattern and climate change as a long-term trend. While vortex events influence surface temperatures, they are not directly linked to climate change. Short-term cold events should not be confused with long-term climate shifts. | Logical_relation: supports`	`supports. The abstract explains how the polar vortex affects winter circulation and temperature anomalies. Cold signatures may reflect natural variability rather than contradicting climate change, thus supporting the claim.`
Mistral-Nemo (Zero-shot)	`The claim aligns with the data. The abstract confirms that the polar vortex is a natural weather pattern and that climate change is a long-term trend, separate from short-term weather events. It also notes that the polar vortex can influence winter weather, but its effects are diverse and not always associated with large-scale cold conditions. Therefore, a cold winter (like those caused by the polar vortex) does not refute the reality of long-term climate change. | Relation: supports`	`supports. Reason: Weak vortex events cause cold-air outbreaks as part of natural variability. These do not contradict long-term warming trends. Summary: Cold events are natural variability and do not refute climate change.`
T5-base (Fine-tuned)	`The abstract identifies the polar vortex as a key driver of winter variability and associates weak and strong vortex winters with climate anomalies. This supports the claim that cold conditions do not negate long-term climate change. | Relation: Supports`	`Relation: Supports | Rationales: The polar vortex drives winter variability and climate anomalies. Cold temperatures therefore, reflect natural variability rather than contradicting long-term climate change.`

Table 6: Zero-shot prompting examples (GPT-4o-mini, Phi-3 Medium, Mistral-Nemo) are shown in the first three rows. Fine-tuned T5-base is shown in the last row. Left column presents intermediate reasoning or explanation→prediction; right column presents post-hoc rationalization or prediction→explanation.

4. Related Work

Large language models (LLMs) have demonstrated remarkable capabilities in reasoning and complex inference tasks due to their massive scale and extensive pre-training on diverse corpora. In particular, Chain-of-Thought (CoT) prompting (Wei et al., 2022) enables models to generate step-by-step intermediate reasoning, improving performance on tasks that require multi-step deduction. Zero-shot CoT prompting (Kojima et al., 2022), using simple instructions such as *Let's think step by step*, further shows that LLMs can perform structured reasoning even without explicit exemplars. However, performance can vary with task complexity and reasoning type (Huang and Chang, 2023). Prior work

in biomedical claim verification demonstrates the potential of self-explainable inference, where models jointly produce entailment predictions and supporting evidence (Jullien et al., 2023). Liang and Sonntag (2025a) showed that multi-step reasoning prompts can improve LLM performance on complex biomedical claim verification tasks, but these methods are often computationally expensive and yield only minimal gains. Drawing inspiration from biomedical and general scientific claim verification, we adapt these LLM reasoning and retrieval strategies to the climate domain.

5. Conclusion

Overall, we explore climate-related claim verification using LLMs of different sizes under two self-explainable paradigms: intermediate reasoning and post-hoc rationalization. We have conducted experiments comparing the performance of the models, and T5-base fine-tuned with augmented post-hoc rationalizations. Several key findings emerge from the experimental results: model size remains the primary factor influencing reasoning capability and verification performance. Meanwhile, smaller models like T5-base may benefit from fine-tuning with augmented explanations, and post-hoc rationalization provides the greatest gains for these models. There is a clear trade-off between performance, computational efficiency, and reproducibility. Future work will focus on fine-tuning lightweight LLMs, such as Phi-3 and Mistral, to further investigate the efficiency and effectiveness of self-explainable claim verification systems.

Acknowledgments

This work was funded by the German Federal Ministry of Education and Research (BMBF) under grant numbers 01IW24006 (NoIDLEChatGPT), as well as by the Endowed Chair of Applied AI at the University of Oldenburg. We also appreciate the support of a grant from Accenture Labs. We also gratefully acknowledge the support provided by a grant from Accenture Labs[4].

Bibliographical References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Raia Abu Ahmad, Max Upravitelev, Aida Usmanova, Veronika Solopova, and Georg Rehm. 2026. ClimateCheck 2026: Scientific Fact-Checking and Disinformation Narrative Classification of Climate-related Claims. In *Proceedings of the 3rd International Workshop on Natural Scientific Language Processing (NSLP 2026)*, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Raia Abu Ahmad, Aida Usmanova, and Georg Rehm. 2025. The climatecheck dataset: Mapping social media claims about climate change to corresponding scholarly articles. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 42–56.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'regan, Donal Landers, and André Freitas. 2023. SemEval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages

[4]https://iml.dfki.de/news/autoprompt-aims-to-improve-chatgpts-analysis-of-clinical-data/

2216–2226, Toronto, Canada. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 22199–22213.

Siting Liang and Daniel Sonntag. 2025a. Advancing biomedical claim verification by using large language models with better structured prompting strategies. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 148–166, Viena, Austria. Association for Computational Linguistics.

Siting Liang and Daniel Sonntag. 2025b. Explainable Biomedical Claim Verification with Large Language Models. In *Joint Proceedings of the ACM IUI Workshops 2025*, volume 3957 of *CEUR Workshop Proceedings*, Cagliari, Italy. CEUR-WS.org. Co-located with the 30th Annual ACM Conference on Intelligent User Interfaces (IUI 2025).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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.

Yash Saxena, Sarthak Chopra, and Arunendra Mani Tripathi. 2024. Evaluating consistency and reasoning capabilities of large language models.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yu Xia, Rui Wang, Xu Liu, Mingyan Li, Tong Yu, Xiang Chen, Julian McAuley, and Shuai Li. 2024. Beyond chain-of-thought: A survey of chain-of-x paradigms for llms.

Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. 2023. Towards better chain-of-thought prompting strategies: A survey.