The ClimateCheck Dataset: Mapping Social Media Claims About Climate Change to Corresponding Scholarly Articles

Raia Abu Ahmad[1,2], **Aida Usmanova**[3], **Georg Rehm**[1,4]

[1]Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Berlin, Germany
[2]Technische Universität Berlin, Germany [3]Leuphana Universität Lüneburg, Germany
[4]Humboldt-Universität zu Berlin, Germany

Corresponding author: raia.abu_ahmad@dfki.de

Abstract

The rapid spread of misinformation on and through social media poses a significant challenge to public understanding of climate change and evidence-based policymaking. While natural language processing techniques have been used to analyse online discourse on climate change, no existing resources link social media claims to scientific literature. Thus, we introduce ClimateCheck, a human-annotated dataset that connects 435 unique, climate-related English claims in lay language to scientific abstracts. Each claim is connected to at least one and at most seventeen abstracts, resulting in 3,048 annotated claim-abstract pairs. The dataset aims to facilitate fact-checking and claim verification by leveraging scholarly document processing to improve access to scientific evidence in online discussions about climate change.

1 Introduction

Social media serves as a powerful tool to discuss critical issues such as climate change. However, it also accelerates the spread of misinformation (Fownes et al., 2018; Al-Rawi et al., 2021), making it increasingly difficult to ensure an informed public and create evidence-based policies.

Natural language processing techniques have proven valuable in analysing online discourse on pressing topics (Stede and Patz, 2021). A particularly promising application is linking social media discussions to peer-reviewed scholarly articles, fostering an evidence-based public dialogue (Sarrouti et al., 2021). However, previous efforts have primarily focused on the biomedical domain (Saakyan et al., 2021; Mohr et al., 2022) and, to the best of our knowledge, no resources have been developed to facilitate this connection for climate change discourse among the public.

To address this, we introduce **ClimateCheck**, a human-annotated dataset that links atomic English claims in lay language to scientific abstracts related to climate change. Our work aims to support fact-checking efforts and promote scientifically grounded discussions on climate change.

This paper describes the detailed process of developing the ClimateCheck dataset, which consists of four main stages, illustrated in Figure 1: (1) Collection of claims, (2) Collection of publications, (3) Linking claims to abstracts, and (4) Manual annotation of claim-abstract pairs.

We collected claims from several existing sources and decomposed them into an atomic, scientifically check-worthy form. Claims were either directly extracted from social media or synthetically generated using text style transfer techniques. We then sourced abstracts from scholarly articles in climate change and environmental sciences using existing research registries. To efficiently link claims and abstracts, we employed a pooling strategy popularised by TREC (Voorhees, 2005; Harman, 2011), as seen in similar datasets (Wadden et al., 2022). The relevant abstracts for each claim were first retrieved via a sparse retrieval method, followed by a neural cross-encoder for re-ranking. Then, state-of-the-art models identified abstracts containing supporting or refuting evidence. This resulted in claim-abstract pairs manually annotated by five graduate students in climate sciences. We adopt an existing annotation scheme (Thorne et al., 2018a), where each pair is annotated as *supports*, *refutes*, or *not enough information (NEI)*.

This process resulted in 1,325 unique English claims, of which we employ 435 for running the ClimateCheck shared task (Abu Ahmad et al., 2025). We split the data into training and testing sets with 259 and 176 unique claims, respectively. Each claim in the training data is linked to at least one and at most five abstracts based on our own linking approach, while for the testing data, we annotate additional claim-abstract pairs based on the submissions of participants, result-

Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 42–56

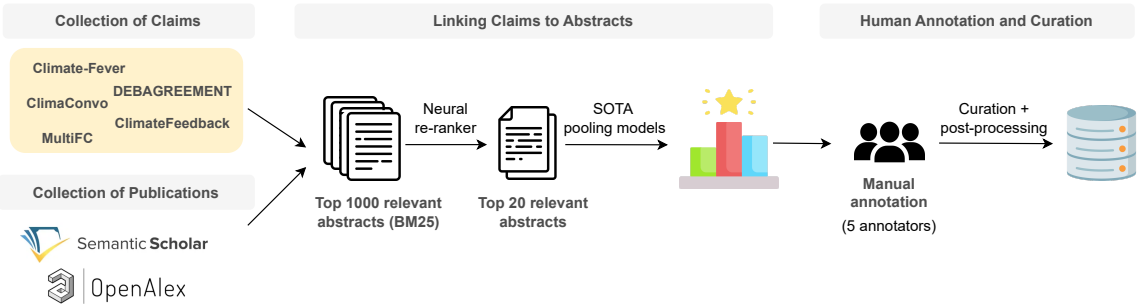

Figure 1: Process of developing the ClimateCheck dataset consisting of four main steps: (1) Collection of claims, (2) Collection of publications, (3) Linking claims to abstracts, and (4) Manual annotation of claim-abstract pairs.

ing in a maximum number of seventeen connected abstracts per claim. The overall process results in **3,048** claim-abstract pairs manually annotated with a total inter-annotator agreement (IAA) score of 0.69 using Cohen's κ.

The rest of the paper is structured as follows. Section 2 reviews existing datasets for scientific fact-checking in general, mentioning those specific to the climate sciences domain. Sections 3, 4, 5, and 6 respectively explain the processes of collecting claims, collecting publications, linking them, and annotating claim-abstract pairs. Importantly, each process is followed by either manual or automatic evaluation to ensure that no errors propagate to the next step.[1] Lastly, Section 7 presents the performance of pooling models on the annotated data and Section 8 concludes our paper.

2 Related Work

Several datasets have been developed for fact-checking across various domains. Fact Extraction and VERification (FEVER, Thorne et al., 2018b) is a benchmark that established a baseline for evidence-based claim verification using Wikipedia as a corpus. Likewise, Wikipedia Citation Entailment (WiCE, Kamoi et al., 2023) is a dataset of fine-grained natural claims mapped to Wikipedia articles that also provides entailment judgements for each subclaim. Although FEVER and WiCE cover a broad range of general knowledge claims, their evidence pool lacks the depth of scientific expertise required to verify domain-specific claims. Similarly, X-Fact (Gupta and Srikumar, 2021) is a multilingual benchmark designed for fact-checking general claims across 25 languages, but does not focus on scholarly publications as evidence sources.

Several fact-checking datasets use scientific literature as an evidence source. For example, SciFact (Wadden et al., 2020) is a benchmark for scientific claim verification in which claims are fact-checked against peer-reviewed biomedical abstracts, with claims annotated as supported or refuted based on evidence sentences. SciFact-Open (Wadden et al., 2022) extends this to an open-domain setting, requiring the retrieval of relevant abstracts before verification. However, a key distinction from our approach is that SciFact and SciFact-Open derive claims from scientific documents rather than public discourse. Closer comparisons to our work are HealthVer (Sarrouti et al., 2021), COVID-Fact (Saakyan et al., 2021), and CoVERT (Mohr et al., 2022), which link social media claims to scientific publications. However, all three are limited to the biomedical domain.

When it comes to climate-related fact-checking efforts, Climate-FEVER (Diggelmann et al., 2020) is, to our knowledge, the only publicly available dataset designed to assess the veracity of claims about climate change. It follows a FEVER-like approach, where claims sourced from English news articles are linked to evidential sentences from Wikipedia. While Climate-FEVER is a valuable resource, it does not align directly with our goal of connecting public discourse in lay language to scientific publications. Other climate-focused fact-checking efforts (Leippold et al., 2024; Augenstein et al., 2019) rely on claims extracted from dedicated fact-checking websites such as Science Feedback[2] and Skeptical Science.[3]

[1] https://github.com/ryabhmd/climatecheck

[2] https://science.feedback.org

[3] https://skepticalscience.com

3 Collection of Claims

To collect claims about climate change, we used five existing sources: (1) Climate-Fever (Diggelmann et al., 2020), (2) DEBAGREEMENT (Pougué-Biyong et al., 2021), (3) ClimaConvo (Shiwakoti et al., 2024), (4) ClimateFeedback,[4] and (5) MultiFC (Augenstein et al., 2019). Some of these directly extract text from social media, while others utilise claims from other sources, such as news outlets. In order to match the claims with the purpose of this dataset, text that was extracted directly from social media platforms underwent a process of claim detection (Section 3.1) and atomic claim generation (Section 3.2), while text extracted from other sources was converted to social media text style using text style transfer techniques (Section 3.3).

3.1 Scientific Claim Detection

Some of the reused datasets were not developed explicitly for fact-checking, thus, raw text did not necessarily contain claims. These include ClimaConvo, a dataset of tweets on climate change originally annotated for relevance, stance, hate speech, and humour (Shiwakoti et al., 2024), as well as DEBAGREEMENT (Pougué-Biyong et al., 2021), a dataset gathered from Reddit, which includes submissions and posts from January 2015 to May 2021 on r/climatechange.[5]

To filter these datasets, we first used an environmental claim detection model fine-tuned on ClimateBERT (Stammbach et al., 2023) to obtain an initial list of potential claims. Then, we manually reviewed the text classified as claims, as well as the text classified as non-claims with a probability of less than 80%. This was done to ensure that we get the maximum number of claims possible from the datasets, without missing any false negatives produced by the claim detection model.

Since the two aforementioned datasets are general public discussions, we noticed that some claims did not refer to scientific topics, rather discussing current political news about climate change. To detect those, we utilised the gemini-1.5-flash model (Gemini Team et al., 2023) in a zero-shot setting, with the self-ask (Press et al., 2023) and rephrase-and-respond (Deng et al., 2023) prompting methods (see Appendix A). This model was selected due to its open availability, fast response time, and competitive performance relative to other models. The model was asked to return a confidence percentage along with its prediction, of which we manually reviewed claims with at least 90% confidence. If any doubt was encountered in terms of the scientific check-worthiness of a claim, it was kept in the dataset, aiming for the climate sciences annotators to decide during the annotation phase of the project.

3.2 Atomic Claim Generation

Fact-checking tasks usually decompose texts to *atomic claims*, which are defined as statements that convey a single, clear, indivisible, and context-independent proposition or piece of information that can be evaluated as true or false (Zhang et al., 2024). More specifically, a *scientific atomic claim* is defined as a statement expressing a finding about one aspect of a scientific entity or process, which can be verified from a single source (Wadden et al., 2020).

Since tweets and posts on Reddit can sometimes contain several atomic claims, we processed them using the gemini-1.5-flash model to extract single atomic claims (see Appendix A). The results were manually reviewed and refined by two near-native English speakers. The instructions given to the refinement process consisted of: (1) Check that the claim indeed exists in the original text; (2) Check that the original text contains the same claim with minimal edits, preserving the original linguistic style; and (3) Check that the claim is atomic using the aforementioned definition. If not, the claim was rephrased to an atomic form when possible. The allowed alterations were replacing pronouns with nouns, adding a subject or an object to elucidate the context, and/or splitting a conjunctional sentence into several atomic claims. Table 7 in Appendix B shows an example of a tweet with several scientific atomic claims, followed by the model output and manual alterations.

3.3 Text Style Transfer

A key objective of this work is to bridge public discourse and scientific knowledge. To that end, we aimed to collect claims that not only reflect discourse on climate change, but also follow the linguistic style in which public conversations usually occur: informal and using colloquial language such as slang, abbreviations, and unconventional grammar (Benamara et al., 2018; Pavlick and Tetreault, 2016). Prior work has shown that such

[4] https://science.feedback.org/climate-feedback
[5] https://www.reddit.com/r/climatechange/

Original Claim	Synthetic Claim
Both the extent and thickness of Arctic sea ice has declined rapidly over the last several decades.	The Arctic sea ice is in trouble! It's been shrinking rapidly in both size and thickness. We gotta do something to turn this around! #SeaIce #ClimateChange

Table 1: Sample of an original claim from a news source and its synthetic tweet-like output.

specialised language requires appropriate datasets and models (Antypas et al., 2023; Barbieri et al., 2022), and recently, Cao et al. (2025) demonstrated that retrieval models underperform when queries are written informally, proving the importance of linguistic registers in datasets used to train and/or fine-tune models.

To be able to reuse datasets that were not developed from social media sources, we rephrased claims to resemble language typically used on social media, inspired by recent research on using large language models (LLMs) for text style transfer (Mukherjee et al., 2024). We used the gemini-1.5-flash model and various prompting techniques, such as role prompting (Schulhoff et al., 2024), rephrase-and-respond, self-ask, and external attention prompting (EAP, Chang et al., 2024) to generate three tweet-style rephrasings per claim (see Appendix A). We then evaluated each rephrased claim based on: (1) BERTScore for similarity to original claim, (2) GPT-2-based perplexity for fluency, and (3) Style classification confidence for text style. These specific evaluation metrics were chosen based on a recent survey on text style transfer (Mukherjee et al., 2024). To develop a style classifier, we fine-tuned BERT-base (Devlin et al., 2019) on a dataset of social media vs. non-social media text. To avoid a classification based on topic rather than style, the texts in both categories dealt with the climate domain. Non-social media sentences were gathered from scientific abstracts, Wikipedia articles, and IPCC reports,[6] and social media texts were taken from the ClimaConvo and DEBAGREEMENT datasets.[7]

To choose a tweet-style representative for each claim, we selected the highest scoring rephrasing as the final output. Tables 1 and 2 present a sample of a rephrased claim and the evaluation averages for all claims, respectively. The results suggest that the rephrased claims are fluent, semantically similar to the original claims, and stylistically similar to social media rather than formal text.

Metric	Score
Perplexity	34.54
BERTScore	72.93
Class. prob. of "social-media" class	99.87

Table 2: Average evaluation scores of the chosen synthetic tweets.

The processes described above resulted in **1,325** English claims. Table 3 summarises the claims by source and original style, and Figure 2 presents four claim samples from various sources.

To illustrate topic diversity in the final set of claims, we ran BERTopic (Grootendorst, 2022) and grouped the results into 16 clusters based on keywords and representative documents. These clusters were then reviewed and named manually, the results of which are shown in Figure 3. To connect the clusters with existing work, we looked for representative (sub-)topics in climate change. However, existing topic lists and taxonomies are usually developed based on official documents and reports rather than public discourse (Sica et al., 2023), or focus on misinformation in discourse rather than presenting neutral topics (Coan et al., 2021). That being said, we manually mapped our resulting clusters to the topics of the World Data Center for Climate (WDCC),[8] as well as the taxonomy presented by Sica et al. (2023). The results are shown in Appendix B.

4 Collection of Publications

To build the corpus of scholarly publications, we first queried S2ORC (Lo et al., 2020) via its bulk search API using "climate change" as the search term and filtered the results to the Environmental Sciences field, yielding 210,237 publications. To better simulate a real-life fact-checking environment with millions of available studies, we expanded the corpus using OpenAlex (Priem et al., 2022), filtering for open-access English publications on climate change, which yielded 826,531

[6] https://www.ipcc.ch/data/

[7] The classification model is available at https://huggingface.co/rabuahmad/cc-tweets-classifier

[8] https://www.wdc-climate.de/ui/topics

Dataset	Source	Original Style	No. of Claims
Climate-Fever	News Articles	Formal	741
DEBAGREEMENT	Reddit	Informal	274
ClimaConvo	Twitter	Informal	164
ClimateFeedback	Media	Formal	97
MultiFC	Diverse	Formal	49
Total			**1325**

Table 3: Overview of datasets reused to collect climate change-related claims.

Source: ClimaConvo
Generation method: Original
Claim: "burning bioenergy accelerates climate change"

Source: DEBAGREEMENT
Generation method: Original
Claim: "Organic farming to build up organic matter in soil can sequester large amounts of carbon from the atmosphere"

Source: ClimateFeedback
Generation method: Synthetic
Claim: "Apparently, changes in Earth's orbit and tilt, not human activity, are responsible for global warming. 🤔"

Source: Climate-Fever
Generation method: Synthetic
Claim: "Scientists used to think the Arctic would be ice-free in summer by 2013. #GlobalWarming"

Figure 2: Samples of claims in the dataset.

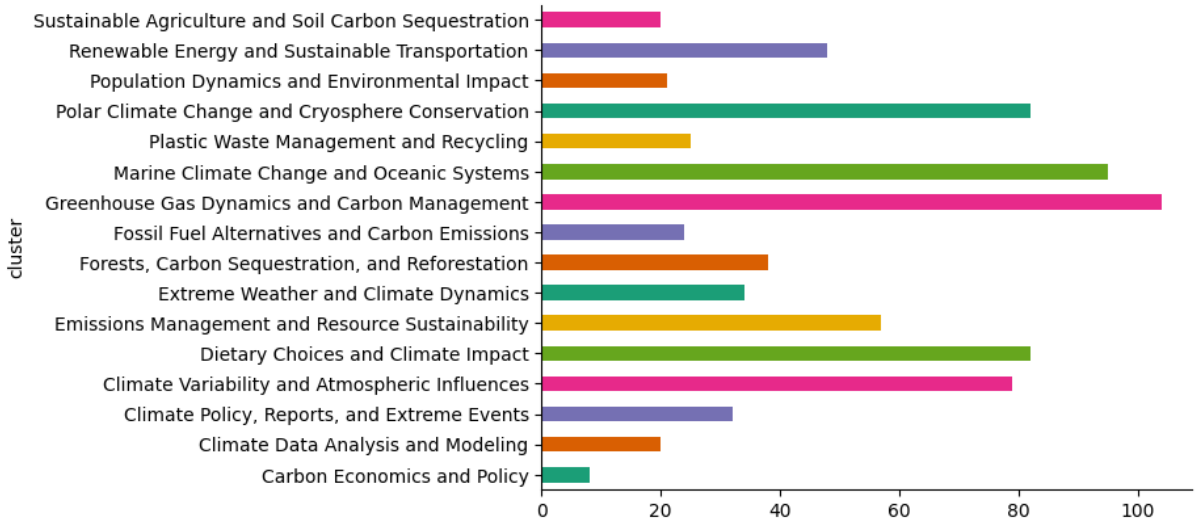

Figure 3: Distribution of represented topics in the collected claims; produced automatically using BERTopic.

articles. After merging and deduplicating by DOI, we further filtered out non-English[9] publications and those missing abstracts or full-text URLs, resulting in a corpus of 835,659 publications. Upon inspection, we noticed that some publications in the corpus were noisy, consisting of think-pieces and various non-peer-reviewed documents. Thus, we filtered out publications with less than 10 citations as a quality measure, chosen based on similar prior research (Wadden et al., 2022). The final corpus consists of **394,269** publications.[10]

5 Linking Claims to Publications

Following retrieval strategies popularised by the TREC competitions (Voorhees, 2005; Harman, 2011; Wadden et al., 2022), we linked each claim to relevant abstracts of scholarly articles using a sparse retrieval method, followed by a neural re-ranker. We then employed a pooling approach, using six state-of-the-art models to classify each

[9]Using https://github.com/fedelopez77/langdetect

[10]The publications corpus is available at https://huggingface.co/datasets/rabuahmad/climatecheck_publications_corpus

claim-abstract pair as "Supports", "Refutes", or "NEI". If an abstract was classified as evidentiary (i. e., either supports or refutes) by at least three models, the claim-abstract pair was added to the annotation pool.

For the sparse retrieval step, we used BM25 (Robertson et al., 2009), a method that relies on TF-IDF keyword matching, to get the top 1000 abstracts per claim. Then, we used a BERT-based neural cross-encoder[11] trained on the MS MARCO passage ranking task[12] to re-rank the retrieved abstracts per claim. For the pooling step, we chose six models based on the following criteria: (1) Open source, (2) Available on Hugging Face for ease of implementation, (3) Parameter size falls between 120M and 15B due to a limit in compute resources, (4) State-of-the-art performance on language understanding, natural language inference (NLI), and/or claim verification tasks. We checked the last criterion using the SuperGLUE,[13] OpenLLM,[14] and MTEB[15] leaderboards.

Consequently, three sequence classification models and three causal LLMs were chosen: 1. RoBERTa-large, fine-tuned on the MNLI dataset,[16] 2. DeBERTa-xxlarge, fine-tuned on the MNLI dataset,[17] 3. XLM-RoBERTa, fine-tuned on the XNLI dataset,[18] 4. Yi-1.5-Chat with a 16K context window (Young et al., 2024), 5. Qwen 1.5-14B-Chat (Bai et al., 2023), and 6. Llama3.1 8B-Instruct.[19] Due to compute and time limitations, the top 20 abstracts from the re-ranking phase were considered. Causal models were prompted using zero-shot role prompting and chain-of-thought (Wei et al., 2022) techniques (see prompt in Appendix A). We used HuggingFace's pipeline object[20] with the default text generation hyperparameters, disabling sampling, thus effectively selecting the most likely next token at each step. For sequence classification models, we used the default forward pass without any adjustments.

To prepare the annotation corpus, a maximum of five abstracts per claim were considered, preferring higher-ranking evidentiary abstracts. Interestingly, the output of the linking phase resulted in a total of **1,167** unique claims connected to a minimum of 1 and a maximum of 5 abstracts. Hence, 158 claims were naturally filtered out due to not resulting in any connected abstracts in the pooling process. These were indeed non-scientifically check-worthy claims that were mistakenly left in the data during the claim collection process (see examples in Appendix B).

6 Annotation Process

To annotate the corpus of claim-abstract pairs, we hired five part-time graduate students (master's). All students have strong expertise in climate sciences, as evidenced by their academic records, and are enrolled in English-language programmes dealing with different aspects of climate sciences. Their English proficiency was proven by providing official results from certified English language tests. We used the INCEpTION (Klie et al., 2018) annotation tool, which offers an automatic calculation of IAA and allows multiple users and roles (Borisova et al., 2024).

The annotation process followed these steps: (1) Read the claim carefully. (2) Read the abstract carefully. (3) Label the pair as one of the following: **"Supports"**: If the abstract supports the claim. **"Refutes"**: If the abstract refutes the claim. **"NEI"**: If the abstract does not provide sufficient information. Annotators were explicitly asked to decide only based on the given abstract, not on their prior knowledge.

To account for mistakes in the data preparation process in terms of creating atomic claims, the annotators were asked to report cases that were manually reviewed. If a claim was shortened to an atomic form, both annotators were updated and asked to annotate with the new version of the claim. Additionally, if an annotator encountered a claim that is not check-worthy against scientific articles, it was disregarded.[21]

Due to time and resource restrictions, the first version of the dataset contains a total of 435

[11] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2

[12] https://github.com/microsoft/MSMARCO-Passage-Ranking

[13] https://gluebenchmark.com/leaderboard

[14] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/

[15] https://huggingface.co/spaces/mteb/leaderboard

[16] https://huggingface.co/FacebookAI/roberta-large-mnli

[17] https://huggingface.co/microsoft/deberta-v2-xxlarge-mnli

[18] https://huggingface.co/joeddav/xlm-roberta-large-xnli

[19] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[20] https://huggingface.co/docs/transformers/en/main_classes/pipelines

[21] Full guidelines given to annotators are available at: https://github.com/ryabhmd/climatecheck/blob/master/ClimateCheck%20Annotation%20Guidelines.pdf

unique claims resulting in 1,815 annotated claim-abstract pairs. We split those into training and testing sets, where the former includes 259 unique claims and 1,144 claim-abstract pairs, while the latter 176 unique claims with 671 claim-abstract pairs. Each document was annotated by two students and curated by a third in case of disagreement. For administrative reasons, we had two annotation groups for the training data, and three for the testing data, each group consisting of two annotators given the same documents.

The dataset was used for the ClimateCheck shared task (Abu Ahmad et al., 2025), where we annotated a subset of claim-abstract pairs from the submissions of participants on a weekly basis, using claims from the test set. This resulted in 1,233 additional manually annotated claim-abstract pairs, with a total of **3,047** documents overall. Figure 4 shows the number of claims as a function of the number of connected abstracts, and Table 4 displays the distribution of labels in each split.[22]

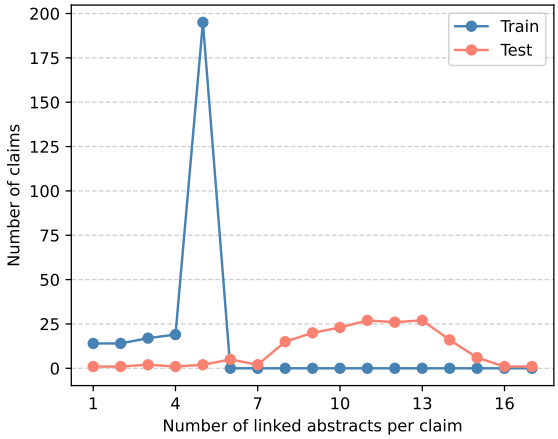

Figure 4: Distribution of the number of abstracts connected to unique claims in the train and test splits.

	Train	Test	Total
Supports	446	749	1195
Refutes	241	266	507
NEI	457	889	1346
Total	1144	1904	3048

Table 4: Label distribution in the train and test splits of the dataset.

The quality of annotations is evaluated based on IAA using Cohen's κ for pairwise comparisons (Cohen, 1960). While this metric was introduced to account for chance agreement, interpretive and widely-used guidelines, such as those by Landis and Koch (1977), suggest that values between 0.61 and 0.80 indicate substantial agreement. Our annotation process achieved an overall Cohen's κ score of 0.69, suggesting a high level of consistency among annotators. Throughout the project, special attention was paid to the agreement score, with the curator flagging claims with low IAA across all associated abstract pairs as potentially vague. Those were then reviewed and rephrased when necessary. The final IAA results are shown in Table 5, indicating individual group agreements in each data split. Importantly, the overall scores are weighted averages, taking the number of annotated documents into account.

Group	IAA	# of Documents
Training Data		
Group 1	0.74	607
Group 2	0.71	537
Overall	0.73	1144
Testing Data		
Group A	0.68	570
Group B	0.62	576
Group C	0.68	758
Overall	0.66	1904
Total	**0.69**	**3048**

Table 5: IAA results for annotated claim-abstract pairs measured using Cohen's κ.

7 Performance of Pooling Models

After finalising the annotations for both the train and test sets, we evaluated the pooling models on the task of claim verification using all annotated documents. The results are shown in Table 6, reported using weighted scores of precision (P), recall (R) and F1, as well as accuracy (Acc.). The sequence classification models were not fine-tuned on the dataset, and the causal models were prompted in a zero-shot setting (see Appendix A).

We note that the sequence classification models achieve similar levels of performance, ranging between an F1 score of 0.31 - 0.33 and an accuracy score of 0.43, with an overwhelming bias toward predicting the NEI class. The results of these models indicate frequent misclassifications of true evidentiary classes, indicating a limitation in models fine-tuned on general NLI datasets, such as MNLI

[22]The ClimateCheck dataset is publicly available at https://huggingface.co/datasets/rabuahmad/climatecheck

and XNLI, when it comes to their applicability to more domain-specific data. We hypothesise that the climate jargon in claims, technical terminology in abstracts, and the overall complex causal structures in claim-abstract pairs is not well represented in standard benchmarks, further supporting the need for domain- and register-specific datasets like ClimateCheck.

In contrast, instruction-tuned LLMs show a significantly better performance, with Yi-1.5-9B-Chat-16K achieving the best F1 and accuracy scores, both 0.61. This suggests that such models have more generalised reasoning abilities and contextual understanding, likely due to their exposure to such data during training. Interestingly, Yi outperformed Qwen by a large margin, despite having fewer parameters, showing that more parameters does not necessarily mean better performance.

Among causal LMs, Yi achieves a relatively balanced precision and recall scores, suggesting that it captures claim-abstract relations with minimal trade-off. However, other models show a clear precision-recall gap, with a pronounced emphasis on weighted precision at the expense of recall. This indicates that while models are highly reliable to make an accurate classification, they miss a significant portion of true instances, generating more false negatives.

Model	P	R	F1	Acc.
roberta-large-mnli	0.40	0.43	0.32	0.43
deberta-v2-xxlarge-mnli	0.39	0.42	0.33	0.43
xlm-roberta-large-xnli	0.40	0.43	0.31	0.43
Yi-1.5-9B-Chat-16K	0.65	**0.61**	**0.61**	**0.61**
Qwen1.5-14B-Chat	0.65	0.53	0.47	0.53
Llama-3.1-8B-Instruct	**0.66**	0.50	0.52	0.50

Table 6: Performance of the six pooling models on the ClimateCheck annotated data using a zero-shot setting.

8 Conclusion

This paper introduces our work of constructing **ClimateCheck**, a human-annotated dataset designed to bridge the gap between social media claims about climate change and corresponding scientific literature. Our dataset consists of 435 unique English claims in lay language, each linked to up to seventeen relevant scientific abstracts, resulting in **3,048** claim-abstract pairs. Claims were fetched from existing resources and refined into atomic, scientifically check-worthy statements, while abstracts were retrieved from open-access climate science publications. We employed BM25 and a neural cross-encoder to rank abstracts per claim, followed by a pooling approach using state-of-the-art models to select the most relevant evidentiary abstracts for annotation. To ensure high-quality annotations, we conducted a structured human annotation process with five graduate students in climate sciences. With this work, our aim is to advance climate-related fact-checking research, fostering a more scientifically grounded public discourse on climate change. Further work can utilise our dataset for tasks such as detecting scientifically check-worthy statements on social media, retrieving relevant publications, and verifying climate-related claims.

Limitations

Although we believe the dataset to be a valuable resource for scientific fact-checking models, it still has several limitations. First, claims are limited to the English language, which hinders improvements in cross-lingual applications that bridge global public discussions with scientific documents. In addition, when linking claims to publications, we only considered abstracts, not the full texts of publications, which might contain more relevant information on a query. During the same step, we filtered abstracts from publications with less than 10 citations as a quality measure, removing informative publications with a smaller citation count. This creates a limitation that could be mitigated in future work by filtering based on other criteria, such as the venue of publication. Additionally, due to time constraints and annotator capacity limitations, we only annotated about a third of the unique claims we originally extracted. However, we plan to release a second version of the dataset with more unique claims in the training data. That being said, we acknowledge that the annotation process is limited in that it is done on a paragraph-level, thus specific sentences that are most informative cannot be connected directly to the claim.

Ethical Statement

The ClimateCheck dataset does not contain sensitive or personal information and is collected from open-source resources. ClimaConvo tweets were preprocessed and thus cannot be traced back to their original form and remain anonymous. Annotators were compensated through a typical pay-

ment scheme and have been informed about the further use of the annotations.

Acknowledgements

This work was supported by the consortium NFDI for Data Science and Artificial Intelligence (NFDI4DS)[23] as part of the non-profit association National Research Data Infrastructure (NFDI e. V.). The consortium is funded by the Federal Republic of Germany and its states through the German Research Foundation (DFG) project NFDI4DS (no. 460234259). We thank the annotators: Emmanuella Asante, Farzaneh Hafezi, Senuri Jayawardena, Shuyue Qu, and Gokul Udayakumar for their work.

References

Raia Abu Ahmad, Aida Usmanova, and Georg Rehm. 2025. The ClimateCheck shared task: Scientific fact-checking of social media claims about climate change. In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria.

Ahmed Al-Rawi, Derrick O'Keefe, Oumar Kane, and Aimé-Jules Bizimana. 2021. Twitter's fake news discourses around climate change and global warming. *Frontiers in Communication*, 6:729818.

Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, Leonardo Neves, Kiamehr Rezaee, Luis Espinosa-Anke, Jiaxin Pei, and Jose Camacho-Collados. 2023. SuperTweetEval: A challenging, unified and heterogeneous benchmark for social media NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12590–12607, Singapore. Association for Computational Linguistics.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. 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 4685–4697.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Farah Benamara, Diana Inkpen, and Maite Taboada. 2018. Introduction to the special issue on language in social media: Exploiting discourse and other contextual information. *Computational Linguistics*, 44(4):663–681.

Ekaterina Borisova, Raia Abu Ahmad, Leyla Garcia-Castro, Ricardo Usbeck, and Georg Rehm. 2024. Surveying the FAIRness of annotation tools: Difficult to find, difficult to reuse. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 29–45, St. Julians, Malta. Association for Computational Linguistics.

Tianyu Cao, Neel Bhandari, Akhila Yerukola, Akari Asai, and Maarten Sap. 2025. Out of style: Rag's fragility to linguistic variation. *arXiv preprint arXiv:2504.08231*.

Yuan Chang, Ziyue Li, and Xiaoqiu Le. 2024. Guiding large language models via external attention prompting for scientific extreme summarization. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 226–242.

Travis G Coan, Constantine Boussalis, John Cook, and Mirjam O Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific reports*, 11(1):22320.

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

Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. 2023. Rephrase and respond: Let large language models ask better questions for themselves. *arXiv preprint arXiv:2311.04205*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

[23] https://www.nfdi4datascience.de

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Jennifer R Fownes, Chao Yu, and Drew B Margolin. 2018. Twitter and climate change. *Sociology Compass*, 12(6):e12587.

Google Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. 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 675–682.

Donna Harman. 2011. *Information retrieval evaluation*. Morgan & Claypool Publishers.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7561–7583. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: System demonstrations*, pages 5–9.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Markus Leippold, Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, et al. 2024. Automated fact-checking of climate change claims with large language models. *arXiv preprint arXiv:2401.12566*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. 2020. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983.

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.

Sourabrata Mukherjee, Mateusz Lango, Zdenek Kasner, and Ondrej Dušek. 2024. A survey of text style transfer: Applications and ethical implications. *arXiv preprint arXiv:2407.16737*.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

John Pougué-Biyong, Valentina Semenova, Alexandre Matton, Rachel Han, Aerin Kim, Renaud Lambiotte, and Doyne Farmer. 2021. Debagreement: A comment-reply dataset for (dis) agreement detection in online debates. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

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.

Mourad Sarrouti, Asma Ben Abacha, Yassine M'rabet, 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.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. 2024. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*.

Shuvam Shiwakoti, Surendrabikram Thapa, Kritesh Rauniyar, Akshyat Shah, Aashish Bhandari, and Usman Naseem. 2024. Analyzing the dynamics of

climate change discourse on twitter: A new annotated corpus and multi-aspect classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 984–994.

Francesco Sica, Francesco Tajani, Mª Paz Sáez-Pérez, and José Marín-Nicolás. 2023. Taxonomy and indicators for esg investments. *Sustainability*, 15(22):15979.

Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066.

Manfred Stede and Ronny Patz. 2021. The climate change debate and natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 8–18.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. 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.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and verification (fever) shared task. *EMNLP 2018*, 80(29,775):1.

EM Voorhees. 2005. Trec: Experiment and evaluation in information retrieval.

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.

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. *arXiv preprint arXiv:2210.13777*.

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.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Zhihao Zhang, Yixing Fan, Ruqing Zhang, and Jiafeng Guo. 2024. A claim decomposition benchmark for long-form answer verification. In *China Conference on Information Retrieval*, pages 41–53. Springer.

A Prompts

Scientific Check-worthiness Prompt

Task: Check-worthiness detection.
Definition: Given a claim, identify if it can be fact-checked **against scientific publications in environmental sciences** to determine their accuracy or truthfulness.
Constraints:
1. Keep in mind that the answer of whether the fact is check-worthy is referring to fact-checking against ***scholarly publications in environmental sciences only and not any other field of science***.
2. For every claim, provide the degree of confidence in the answer you provide. The number should be between 0 and 1 with a higher number indicating higher confidence.
3. Give the output in a json format 'result': 'check-worthy', 'confidence': 0.8
4. Before giving your answer, rewrite the prompt, expand the task at hand, and only then respond.
5. If you have follow-up questions, generate them and then answer them before giving the final output.
Claim: [claim]
Output:

Atomic Claim Generation Prompt

Task: Given an input text, give a list of atomic claims in it. Atomic claims are verifiable statements **expressing a finding about one and only one aspect of a scientific entity or process**, which can be verified from a single source.
Constraints:
1. The output should only split different sentences in the input text so that each sentence contains one claim.
2. ** It is extremely important in this task that the style of the text, including the used words and characters, should not be changed, and the text itself should not be rephrased. Claims should be copy-pasted. **
3. ** Each claim should be self-contained without needing more context. A claim should have a subject, a predicate and an object. If a sentence in the input text needs more context to be understood completely, it should not be included in the list of answers. **
4. Before giving your answer, rewrite the prompt, expand the task at hand, and only then respond.
5. If you have follow-up questions, generate them and then answer them before giving the final output.
6. Give your answers in a list in JSON format.
Examples: [examples]
Input text: [text]
Output:

Text Style Transfer Prompt

Task: Given a claim extracted from a news article, produce a rephrasing as if you are a layperson tweeting about it.
Constraints:
1. Take into account stylistic features of social media text such as use of slang and informal language.
2. Do not overdo your text generations. Keep them plausible enough to believe a human wrote them.
3. Introduce variance in rhetoric and syntactic structures of your tweets. **Not every tweet needs to contain a question.**
4. **Generate tweets in a neutral tone. Do not add irony or satire.**
5.**Keep the scientific claim that is present in the original claim**
6. Give three output options in a JSON format that includes a list of the tweets.
7. Before giving your answer, rewrite the prompt, expand the task at hand, and only then respond.
8. If you have follow-up questions, generate them and then answer them before giving the final output.
Examples of tweets about a similar topic: [examples]
Claim: [claim]
Output:

Pooling Models Prompt

You are an expert claim verification assistant with vast knowledge of climate change, climate science, environmental science, physics, and energy science.
Your task is to check if the Claim is correct according to the Evidence. Generate 'Supports' if the Claim is correct according to the Evidence, or 'Refutes' if the claim is incorrect or cannot be verified. Or 'Not enough information' if you there is not enough information in the evidence to make an informed decision.
Evidence: [abstract]
Claim: [claim]
Provide the final answer in a Python list format.
Let's think step-by-step:

B Additional Samples and Figures

B.1 Atomic Claim Generation

Original Text	Plastics are not only a primary marine pollutant but also a significant driver of the climate crisis. Emissions from plastic production will reach a billion tons per year by 2030, and plastic in the environment releases methane and ethylene in a feedback loop. #FridaysforFuture
Gemini-1.5 Output	['Plastics are not only a primary marine pollutant but also a significant driver of the climate crisis.', 'Emissions from plastic production will reach a billion tons per year by 2030.', 'plastic in the environment releases methane and ethylene in a feedback loop.']
Manual Refinement	['Plastics are a primary marine pollutant.', 'Plastics are a significant driver of the climate crisis.', 'Emissions from plastic production will reach a billion tons per year by 2030.', 'plastic in the environment releases methane in a feedback loop.', plastic in the environment releases ethylene in a feedback loop']

Table 7: Example of processing social media text into atomic claims.

B.2 Filtered Claims

The following list contains ten example claims that were filtered out during the linking process. These claims did not result in any linked abstracts that met our criteria of having at least three evidentiary predictions from the pooling models.

1. So, Benny Peiser has backtracked on his criticism. Interesting... Wonder what made him change his mind?
2. people are trying to dispose of plastics in Uganda by burning
3. Florida needs to step up its game when it comes to business regulations. Ranking 45th out of 50 states is not a good look.
4. The 2016 Future Energy Jobs Act is Illinois' most significant climate legislation.
5. Obama warned the U.S. Coast Guard that global warming is the biggest threat to the military and the world. We gotta take climate change seriously! #ClimateAction #ClimateCrisis
6. Google will run entirely on green energy 24/7 without requiring carbon offsets at all by 2030.
7. Luxury *non-*gas cars need to be celebrated.
8. Carbicrete's process is carbon-negative.
9. They also said the company failed to keep adequate servicing records
10. United Kingdom has a special responsibility to provide moral and political leadership on the climate crisis.

B.3 Topic Distribution of Claims

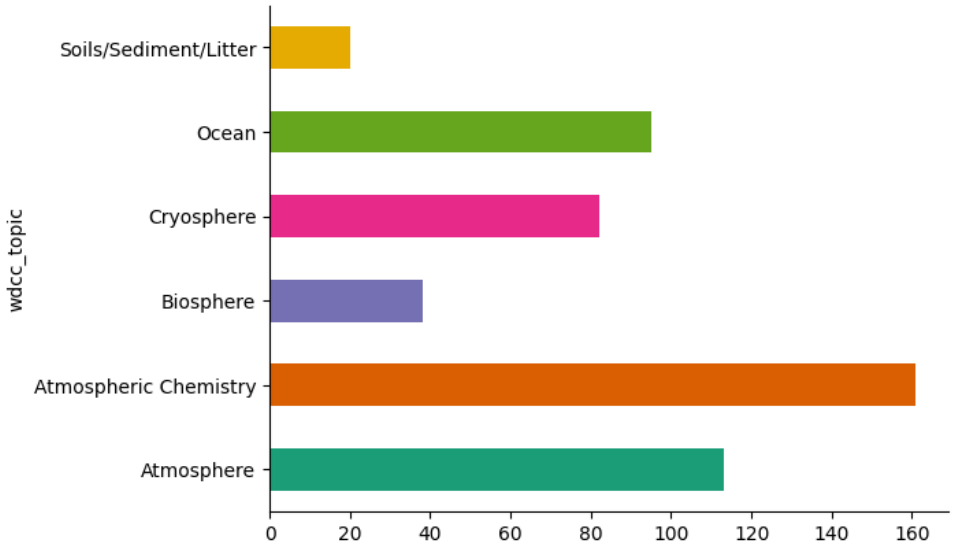

Figure 5: Distribution of represented WDCC topics in the collected claims, made by mapping BERTopic clusters to topics manually.

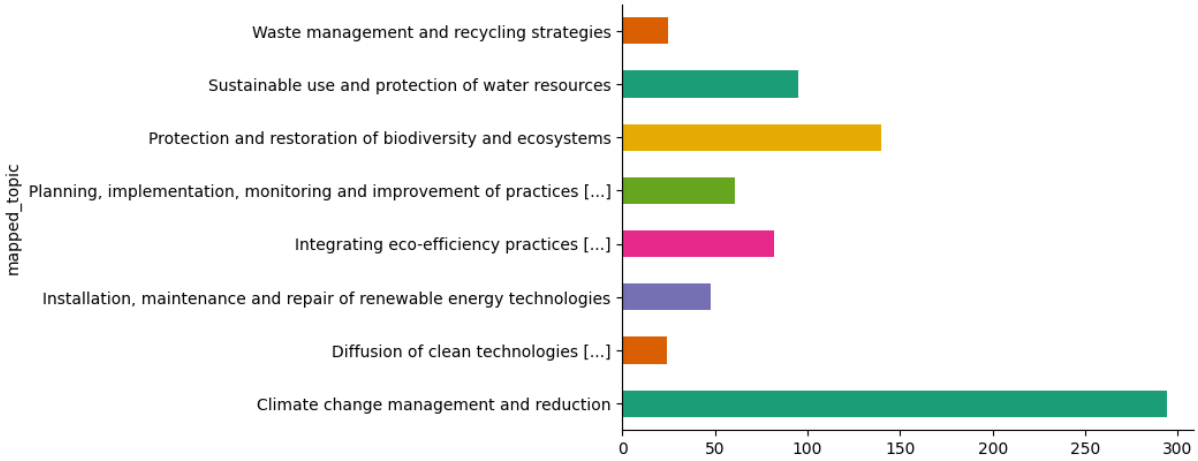

Figure 6: Distribution of claim topics according to the environmental section of the taxonomy presented by Sica et al. (2023).