The ClimateCheck Shared Task: Scientific Fact-Checking of Social Media Claims about Climate Change

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

Misinformation in public discourse on global and significant issues like climate change is often facilitated through social media. However, current systems do not address fact-checking climate-related claims against trustworthy, evidence-based sources, such as scientific publications. To address this, we organised the ClimateCheck shared task at the 5th Scholarly Document Processing (SDP) Workshop, co-located with ACL 2025 in Vienna, Austria. The task featured two subtasks: I. Abstracts retrieval given a claim, and II. Claim verification based on the retrieved abstract. ClimateCheck had 27 registered users with active participation from 13 teams, ten of which submitted results for the first subtask and three for the second. The winning team achieved a Recall@10 score of 0.66 and a Binary Preference score of 0.49 for subtask I, and an F1 score of 0.73 for subtask II. Their method combined sparse retrieval using BM25, an ensemble of fine-tuned cross-encoder models using BGE-rerankers, and LLMs for classification.

1 Introduction

The widespread use of social media has transformed the way people engage with crucial global challenges such as climate change. While these platforms enable a public dialogue, they also fast-track the spread of inaccurate and misleading information (Fownes et al., 2018; Al-Rawi et al., 2021).

Recent work in natural language processing (NLP) offers promising advances in decoding and analysing complex discourse online (Stede and Patz, 2021). Researchers have used methods to detect misinformation (Aldwairi and Alwahedi, 2018; Aïmeur et al., 2023), extract scientific claims and entities (Hafid et al., 2022; Hughes and Song, 2024), and fact-check statements (Guo et al., 2022; Diggelmann et al., 2020). At the same time, work on scholarly document processing has advanced methods for extracting and structuring scientific knowledge (Dagdelen et al., 2024), making it easier to link it to public discourse.

Shared tasks are effective tools for mobilising the research community around challenging tasks, driving innovation and the development of state-of-the-art methods (Filannino and Uzuner, 2018). Previous shared tasks targeted fact-checking by retrieving relevant evidence for a given claim and classifying their relation. However, they mainly focused on non-scientific evidence corpora, e. g., Wikipedia (Thorne et al., 2018; Aly et al., 2021), or were limited to the biomedical domain (Wadden and Lo, 2021). To the best of our knowledge, no previous effort has tackled the challenge of connecting claims posted online about climate change to credible scientific sources.

To address this, we present **the ClimateCheck shared task**, focusing on automatic fact-checking of climate-related claims from social media against scientific publications. The task was hosted at the 5th Scholarly Document Processing (SDP) Workshop[1] and consisted of two subtasks: (I) Retrieving relevant scientific documents for a given claim, and (II) Classifying the claim's veracity based on the retrieved evidence. Subtask I was evaluated using the average scores of Recall@K ($K = 2, 5, 10$) and Binary Preference (Bpref, Buckley and Voorhees, 2004), and subtask II was evaluated using the F1 score in addition to Recall@10 from subtask I.

We used the Codabench platform to host the task (Xu et al., 2022), attracting registrations from 27 users and 13 active teams, ten of which submitted results to the leaderboard.[2] The competition followed the timeline below:

- Training set release: April 1, 2025
- Test set release: April 15, 2025
- Systems submissions deadline: May 16, 2025

[1] https://sdproc.org/2025/

[2] https://www.codabench.org/competitions/6639/

Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 263–275
July 31, 2025 ©2025 Association for Computational Linguistics

- Paper submission deadline: May 23, 2025
- Notification of acceptance: June 13, 2025
- Camera-ready paper due: June 20, 2025
- Workshop date: July 31, 2025

This paper presents an overview of the shared task and summarises the task design (§3), evaluation strategies (§4), dataset preparation (§5), our baselines (§6), approaches of submitted systems (§7), and lessons learned throughout (§8), aiming to inform and encourage future efforts in NLP for mitigating climate change misinformation online.

2 Related Shared Tasks

Several shared tasks have been introduced to support research on automatic evidence retrieval and claim verification. These tasks differ in the domain of claims, the type of evidence corpora, and the complexity of the verification process.

Fact Extraction and VERification (FEVER, Thorne et al., 2018) and its extension, FEVER Over Unstructured and Structured information (FEVEROUS, Aly et al., 2021), were tasks focused on claim verification against Wikipedia articles, the latter expanding into structured evidence such as tables and lists. FEVER established the widely adopted three-stage pipeline of document retrieval, sentence selection, and natural language inference (NLI). However, despite their scale and influence, FEVER and FEVEROUS differ from our effort in their evidence domain, which is encyclopedic rather than scientific, potentially affecting the applicability of certain retrieval methods.

The Automated Verification of Textual Claims (AVeriTeC) shared task was a recent effort presented at the FEVER 2024 Workshop (Schlichtkrull et al., 2024). The task focused on evidence retrieval and veracity prediction of general real-world claims with linked evidence from the web using search engines. This task differs from ours in two main aspects: claims are not domain-specific, and the evidence is retrieved from the web rather than the more trustworthy scientific literature.

The SCIVER shared task was organised at the SDP 2021 workshop, aiming to verify scientific claims extracted from research articles against a given corpus of publications (Wadden and Lo, 2021). Although the task is similar in its focus on scientific evidence, SCIVER's claims originate from research papers and are limited to the biomedical domain, in contrast to our task, which focuses on climate-related claims from public discourse.

Finally, CheckThat!, organised annually as a CLEF lab since 2018 (Nakov et al., 2018), focuses on mitigating misinformation online across different platforms and several languages. Previous editions have addressed claim detection, stance verification, and evidence retrieval, focusing primarily on political and journalistic content. Most recently, the 2025 edition included the Scientific Web Discourse task (Alam et al., 2025), focusing on 1. Detecting whether a post contains references to scientific entities, and 2. Linking posts with implicit references of studies to their relevant publications. These tasks are similar to our work in their objective of connecting public discourse to scientific publications, with task 1 being especially relevant to the pre-processing steps of preparing the ClimateCheck dataset. However, unlike task 2, our work does not assume any mention of a study in a post, rather processing general claims.

3 Task Description

ClimateCheck consisted of two subtasks:

1. **Subtask I – Abstracts Retrieval:** Given a claim from social media about climate change and a corpus of abstracts, retrieve the top 10 most relevant abstracts to the claim.
2. **Subtask II – Claim Verification:** Given the claim-abstract pair received from the previous subtask, classify their relation as 'supports', 'refutes', or 'not enough information (NEI)'.

Participants were allowed to take part either in subtask I only or in both subtasks. The testing dataset consisted of 176 unique claims along with a corpus of 394,269 abstracts from climate-related publications. For the first subtask, the participants were asked to upload a CSV file that includes rows of unique claim-abstract pairs, where each claim was linked to 10 relevant abstracts. If they wished to participate in subtask II, they were asked to add a column denoting the label of the pair. Samples of five claims from the test set along with connected abstracts retrieved by three teams in the competition are available in Appendix A.

4 Evaluation

Subtask I: Abstracts Retrieval

As an information retrieval (IR) task, subtask I tackles identifying relevant pieces of information from large corpora based on a user query. Evaluating IR is an inherently difficult task due to the problem of incomplete relevance annotations when the evidence corpus contains a large number of documents (Buckley and Voorhees, 2004). That is because not all potentially relevant documents can be annotated, making it hard to know whether a system truly failed to retrieve relevant items or simply retrieved items that were never judged.

Various metrics are employed to evaluate IR based on rankings (Buckley and Voorhees, 2004; Järvelin and Kekäläinen, 2002), including Mean Average Precision, Mean Reciprocal Rank, and normalised Discounted Cumulative Gain (Järvelin and Kekäläinen, 2002). However, in our specific task, we faced two primary challenges: the absence of annotated ranking information and the problem of incomplete relevance judgements. Given these constraints, we selected Recall@K and Bpref as our evaluation metrics.

Recall@K measures the proportion of relevant documents retrieved in the top K results. It does not consider the order of the retrieved documents, making it suitable for scenarios where gold ranking information is unavailable. The metric has been widely used to evaluate dense retrieval systems (Karpukhin et al., 2020). In subtask I, we ask participants to retrieve the top 10 abstracts per claim, hence we use $K = 2, 5, 10$ to compare systems on different levels. Bpref is a score designed to handle situations with incomplete relevant judgements. It evaluates how many judged non-relevant documents are retrieved before judged relevant ones, mitigating potential bias introduced by unjudged documents (Buckley and Voorhees, 2004).

The final evaluation of subtask I, which decides the rankings, is the average of the four scores mentioned above. We considered a retrieved abstract to be relevant if it was annotated as evidentiary (i. e., supports or refutes) in our gold data. However, this data was bound to be biased towards our own retrieval method used to create the annotation corpus. Thus, to ensure a fair evaluation, we collected participants' outputs weekly during the test phase, subsequently adding more human-annotated instances to the gold data (see Section 5).

Subtask II: Claim Verification

Claim verification is a classification task, where the system labels each claim-abstract pair retrieved in subtask I as *supports*, *refutes*, or *NEI*, indicating the relation of the abstract to the claim. To evaluate it, we used standard weighted metrics: Precision, Recall, and F1.

Only claim-abstract pairs that have been manually annotated in the gold data were used for evaluation, meaning that unjudged ones were excluded. To ensure a fair comparison across systems, especially since the number of predicted labels varied, the final ranking consisted of the sum of the F1-score and the Recall@10 score from subtask I. This approach rewards systems that not only made accurate classifications, but also retrieved more relevant abstracts, penalising those that have a high F1 score based on only a few examples.

5 Dataset

The foundation of the shared task is the Climate-Check dataset (Abu Ahmad et al., 2025),[3] consisting of 435 unique English climate-related claims in lay language linked to scientific abstracts, resulting in 1,815 claim-abstract pairs. Each pair was reviewed by two graduate students in climate sciences and annotated as *supports*, *refutes*, or *NEI*. In cases of disagreements, a third student curated the claim-abstract pair, deciding its final label.

Claims were collected from available datasets (Diggelmann et al., 2020; Pougué-Biyong et al., 2021; Shiwakoti et al., 2024; Augenstein et al., 2019), and underwent several pre-processing steps: scientific check-worthiness detection, atomic claim generation, and text style transfer, the latter for those not originating directly from social media. The abstracts were collected from OpenAlex (Priem et al., 2022) and S2ORC (Lo et al., 2020), resulting in a corpus of 394,269 climate-related publications.[4] Claims and abstracts were then linked using BM25 (Robertson and Zaragoza, 2009) followed by a cross-encoder trained on the MS-MARCO data and a TREC-like pooling approach using six models to create the annotation corpus. In Abu Ahmad et al. (2025), we describe the development of the dataset in more detail.

The available data was split into training and test-

[3] https://huggingface.co/datasets/rabuahmad/climatecheck

[4] https://huggingface.co/datasets/rabuahmad/climatecheck_publications_corpus

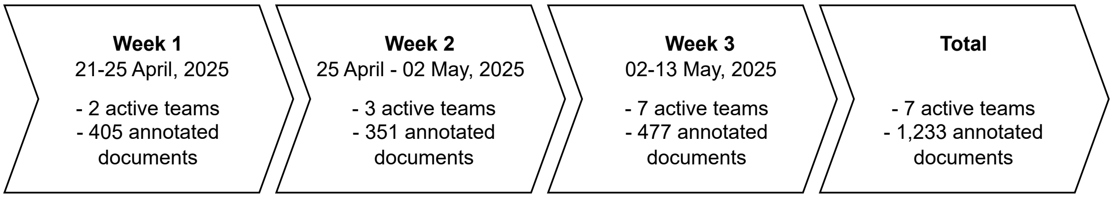


Figure 1: The timeline of our dynamic human annotation process during the testing phase of the ClimateCheck shared task. The process resulted in 1,233 additional claim-abstract pairs added to the gold test data.

ing sets, the former consisting of 259 unique claims and a total of 1,144 claim-abstract pairs, while the latter 176 unique claims and 671 claim-abstract pairs. The annotated pairs of the test set were not released publicly for participants, since they were used as the reference test set for evaluation.

In an attempt to make the evaluation less biased towards the gold test set, which is based on our own linking approach, we annotated more documents on a weekly basis as the task was running. These were based on participants' submissions using the following approach:

1. Every week, we combined the highest-scoring submissions from each active team.
2. For each unique claim-abstract pair, we assessed the agreement among the participating teams (i. e., how many systems retrieved this pair).
3. We annotated pairs with a specific agreement threshold so that as many teams as possible benefit from the new annotations.
4. We updated the gold data with the additional annotations a week later.

The agreement threshold was decided each week depending on the number of submitting teams, taking into account our limited human annotation capacity (four student annotators). If needed, we filtered further based on the rankings of claim-abstract pairs across submitted systems. We summarise the result of this process in Figure 1, and report more details in Appendix B.

To accommodate the timeline of the SDP 2025 workshop and the pace of the annotators, we were able to gather new documents from runs submitted until May 13, 2025, one week before the competition deadline. This process resulted in the addition of 1,233 new claim-abstract pairs added to the gold testing data, with an overall number of 1,904 manually annotated pairs in the gold test set.

6 Baselines

For subtask I, we developed a multi-stage retrieval approach as a baseline, combining sparse and dense retrieval with a neural reranker. BM25 has proven to be a fast and efficient method for initial retrieval (Chen et al., 2017; Nie et al., 2019). We used it as a sparse retrieval step to get an initial set of the top 1000 relevant abstracts per claim. Next, we computed embeddings for each claim and abstract using the msmarco-MiniLM-L-12-v3 sentence transformer,[5] and calculated the cosine similarity for each claim-abstract pair. We selected the top 20 ranked abstracts per claim, filtering out lexically relevant but semantically irrelevant candidates. Finally, a neural reranker, ms-marco-MiniLM-L6-v2,[6] provided cross-encoder scores, resulting in the final candidate pool of the top 10 abstracts per claim.

To obtain labels for each claim-abstract pair as a baseline for subtask II, we used the open source Yi-1.5-9B-Chat-16K model (Young et al., 2024), selected based on our experiments with several models when creating the dataset (Abu Ahmad et al., 2025). The model was prompted in a zero-shot manner with the following 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,
'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.
Only return the verification verdict.
```

[5] https://huggingface.co/sentence-transformers/msmarco-MiniLM-L12-v3

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

7 Submitted Systems and Results

A total of ten teams participated in the ClimateCheck shared task, three of which took part in both abstract retrieval and claim verification tasks. Table 1 summarises the submission statistics and Figure 2 illustrates the amount of submissions throughout the one month testing phase of the task.

Number of registered users	27
Number of active users	13
Number of final submissions (subtask I)	10
Number of final submissions (subtask II)	3
Number of total submissions	613
avg. number of submissions per user	43.64
max. number of submissions by a single user	182

Table 1: ClimateCheck submission statistics.

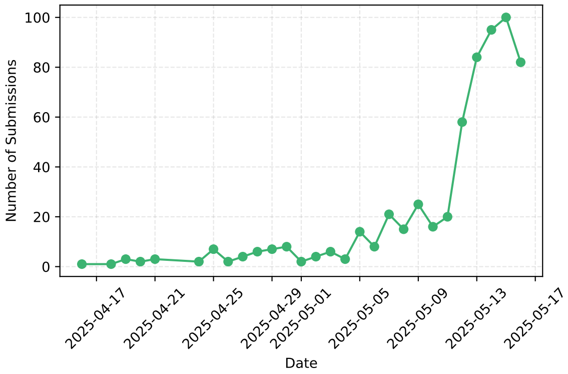


Figure 2: Number of submissions over the one month timeline of the task.

We present the results of subtasks I and II in Tables 2 and 3, respectively. Notably, six teams outperformed our baseline in subtask I, while all of them outperformed it in subtask II. For both subtasks, the winning team is **Ant Bridge**, followed by **akiepura_jlam** in 2nd place, while team **AlexUNLP-FMT** achieves 3rd place in subtask I, and team **EFC** in subtask II. We received system descriptions from the aforementioned top four teams and team **Pranav**, which we briefly summarise below.

7.1 Team Ant Bridge

Team Ant Bridge (Wang et al., 2025) developed a hybrid three-stage approach, combining sparse retrieval, fine-grained reranking, and large language models (LLMs) for claim-abstract classification. As a first step, the team pre-processed all abstract and claim texts to be lowercase, additionally tokenizing and removing punctuation and stopwords. Then, they used BM25 to get the top 5000 abstracts per claim, chosen to maximise recall for the reranking step. In the second stage, they fine-tuned several cross-encoder models based on the BGE-Reranker architecture (Chen et al., 2024a). Training data was constructed as triples of (claim, relevant abstract, irrelevant abstract), with negatives drawn either randomly or as hard negatives, which are abstracts ranked highly by BM25 or semantically close to the claim but not evidentiary. Rerankers were trained using a marginal ranking loss, and their outputs were aggregated using Reciprocal Rank Fusion (RRF, Cormack et al., 2009) to produce the top 10 abstracts per claim.

For subtask II, the team used Gemini 2.5 Pro (Gemini Team et al., 2023) to perform claim-abstract relation classification. Their prompting strategy included persona and task definitions, and supported batch processing of multiple claim-abstract pairs. Additionally, they included distribution guidelines in the prompt to steer the model toward a more balanced output, explicitly instructing it to ensure that the proportion of NEI labels remained at or above 30%. This soft calibration approach helped mitigate bias in label distribution and improved robustness in classification.

7.2 Team akiepura_jlam

The akiepura_jlam team (Kiepura and Lam, 2025) employed a three-stage retrieval and reranking pipeline for subtask I, starting with a hybrid retrieval system that fused BM25, dense and sparse neural retrieval methods using RRF. Their dense model was based on a fine-tuned BGE-M3 encoder (Chen et al., 2024b) trained using triples of (claim, relevant abstract, irrelevant abstract), where NEI-labelled abstracts from the training data served as the negative samples. Dense embeddings were computed for all abstracts, and claim-abstract similarity scores were obtained via dot products. For sparse retrieval, they used SPLADE-v3 (Lassance et al., 2024) to generate high-dimensional vectors for claims and abstracts. The retrieval results from all three methods, BM25, SPLADE, and BGE-M3, were combined with RRF, and the top 600 abstracts per claim were selected for further reranking.

Their second stage comprised of a cross-encoder reranker based on ms-marco-MiniLM-L-6-v2[7] (Wang et al., 2020), which was fine-tuned on the ClimateCheck data using the top 200 candidates

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

Rank	Team	Recall@2	Recall@5	Recall@10	Bpref	Subtask I Score
1	Ant Bridge	*0.21848*	**0.45112**	**0.66476**	**0.49470**	**0.45727**
2	akiepura_jlam	**0.23085**	*0.44128*	*0.60061*	*0.48179*	*0.43863*
3	AlexUNLP-FMT	0.20997	0.39627	0.59112	0.46348	0.41521
4	EFC	0.21769	0.40582	0.57411	0.44952	0.41178
5	gmguarino	0.18064	0.3386	0.47696	0.38678	0.34574
6	Pranav	0.17988	0.31059	0.44038	0.37614	0.32675
–	Our baseline	0.1947	0.30468	0.34359	0.29803	0.28525
7	vanguard	0.1065	0.18062	0.27069	0.243	0.2002
8	nakrayko	0.11499	0.16868	0.27069	0.24266	0.19926
9	lephuquy	0.11101	0.15483	0.15759	0.14953	0.14324
10	seniichev	0.07889	0.12156	0.17622	0.14888	0.13139

Table 2: Results of Subtask I: Abstracts Retrieval; top result in bold, runner-up italicised, third place underlined.

Rank	Team	Precision	Recall	F1	Subtask II Score
1	Ant Bridge	**0.72905**	**0.72644**	**0.72528**	**1.39004**
2	akiepura_jlam	0.69496	0.69726	0.69573	*1.29634*
3	EFC	*0.71676*	*0.71746*	*0.71696*	1.29107
–	Our baseline	0.65448	0.62603	0.63148	0.97507

Table 3: Results of Subtask II: Claim Verification; top result in bold, runner-up italicised, third place underlined.

from Stage 1. Training again involved both positive (evidentiary) and negative (NEI and random) examples. The top 20 reranked abstracts were passed to Stage 3, where a few-shot LLM-based reranker was used, namely RankGPT (Sun et al., 2023) using GPT-4.1[8]. RankGPT treated reranking as a permutation task, reasoning over the full set of abstracts per claim to produce a final ordering. Their final ranking combined the LLM's output with the semantic precision score from the cross-encoder. An ablation study demonstrated the incremental benefits of each stage, showcasing the effectiveness of the entire pipeline.

For subtask II, team akiepura_jlam experimented with both zero- and few-shot prompting, as well as fine-tuned transformer classifiers, with their best performance coming from a hybrid zero-shot prompt that first asked the LLM to determine whether an abstract was evidentiary and if so, to assess whether it supported or refuted the claim.

7.3 Team AlexUNLP-FMT

Team AlexUNLP-FMT (Fathallah et al., 2025) participated only in subtask I, proposing a hybrid retrieval and adaptive reranking strategy to address the limitation of excluding relevant documents in the initial retrieval step. The team combined sparse retrieval, using BM25, with dense retrieval, using a fine-tuned Stella-en-400M-v5 (Zhang et al., 2024) in a contrastive learning approach. From each retrieval method, they extracted the top 50 abstract candidates from the original set of publications. The candidates from both methods were combined, deduplicated, and an initial reranking set was formed. This was followed by the ms-marco-MiniLM-L12-v2 reranker[9] obtaining the top 30 abstracts from the initial reranking set.

For each of the 30 abstracts, the top 10 were selected by choosing the closest neighbours from a similarity graph. The graph was constructed from the entire abstract corpus using the all-MiniLM-L6-v2 bi-encoder model[10]. Each abstract in the graph was connected to the top 10 most semantically similar abstracts, and an iterative process of augmenting candidate sets with semantic neighbours was repeated 20 times. In the last iteration, the top 10 most relevant abstracts with respect to a given claim were selected.

7.4 Team EFC

Similar to other teams, EFC's pipeline included sparse and dense retrieval stages followed by a reranker (Upravitelev et al., 2025). First, 1,500 abstracts were retrieved via BM25, further reduced using a fine-tuned e5-large-v2 model[11] to 150 ab-

[8] https://openai.com/index/gpt-4-1/

[9] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L12-v2

[10] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[11] https://huggingface.co/intfloat/e5-large-v2

Team	Sparse Retrieval	Dense Retrieval	Cross-Encoder	LLM	Graph
Ant Bridge	BM25	✗	BGE-reranker*	✗	✗
akiepura_jlam	BM25, SPLADE	BGE-M3*	MiniLM*	GPT-4.1	✗
AlexUNLP-FMT	BM25	Stella*	MiniLM	✗	MiniLM
EFC	BM25	E5*	MiniLM	✗	✗
Pranav	SPLADE	✗	✗	Gemini-2.0-Flash	✗
Our baseline	BM25	MiniLM	MiniLM	✗	✗

Table 4: Summary of retrieval systems used for subtask I; * indicates fine-tuning with contrastive learning.

Team	LLM	Classification Setup
Ant Bridge	Gemini 2.5	ZS + distribution guidelines
akiepura_jlam	GPT-4.1	Hybrid ZS
EFC	Qwen 14B	ZS w/ reasoning
Our baseline	Yi-1.5-9B-Chat-16K	ZS

Table 5: Summary of classification models used for subtask II (ZS = zero-shot).

stracts. The model was fine-tuned on the entire ClimateCheck training set for three epochs, utilising a contrastive learning approach with positive and negative samples, the latter mined by retrieving the three least relevant publications using their dense retrieval method. Finally, the ms-marco-MiniLM-L12-v2 reranker, also used by Team AlexUNLP-FMT, was applied to get the top 10 relevant abstracts per claim.

To minimise computational inference cost, the team chose to compare smaller encoder-only architectures with larger decoder-only LLMs for subtask II. Their best-performing encoder only model was DeBERTa-v3-large[12], fine-tuned on several NLI datasets as well as the ClimateCheck dataset, while the best LLM was Qwen3 with 14B parameters (Yang et al., 2025). Their best results, those submitted to the leaderboard, were achieved using the Qwen model. However, the team demonstrated that the fine-tuned DeBERTa is not far behind, with a total score of 1.257 in subtask II, while requiring about 0.0026 of the runtime that Qwen needs.

7.5 Team Pranav

Team Pranav participated only in subtask I, utilising a two-stage retrieve-and-rerank approach. They start with sparse retrieval using SPLADE-v3[13] by indexing the entire publications corpus with sparse vector representations. Then, for each claim, they calculate the dot product similarity to retrieve the top 40 abstracts. The second stage of the approach is based on LLM reranking using the Gemini-2.0-Flash[14] model with a list-wise strategy. The LLM is presented with all 40 candidates simultaneously, prompting it to rerank and output the top 10 abstracts that provide evidence to the claim.

8 Discussion

The submissions to ClimateCheck reveal key design patterns and trade-offs in building claim verification pipelines grounded in scientific literature. Although architecture choices varied, several common effective strategies emerged across top-performing teams. We summarise the approaches for subtasks I and II in Tables 4 and 5, respectively, and compare their results visually in Figure 3.

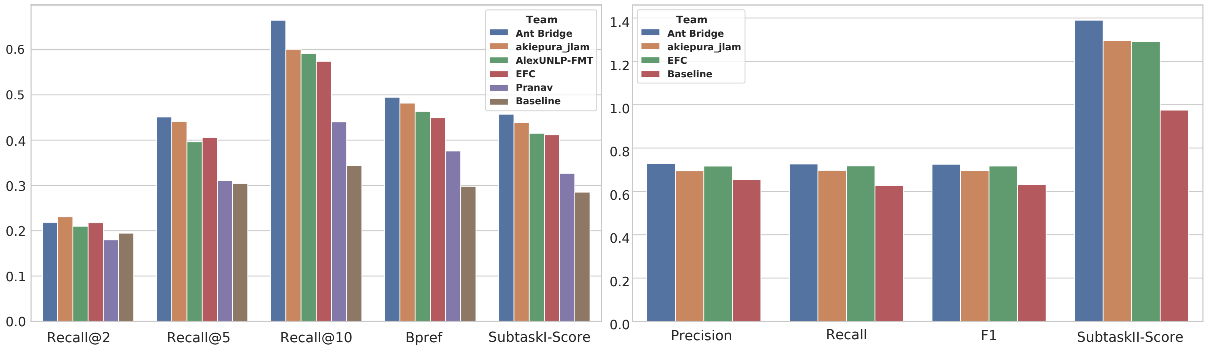
A clear pattern from subtask I is the use of hybrid pipelines, combining sparse retrievers (e.g., BM25 and SPLADE) with different dense retrievers, as well as cross-encoder rerankers (e.g., BGE and MiniLM). Three teams extended this by utilising more advanced components: LLM-based reranking (akiepura_jlam and Pranav) and graph-based reranking (AlexUNLP-FMT). Although the teams achieved competitive scores, they were still outperformed by the relatively simpler ensemble of fine-tuned cross-encoders using RRF presented by Ant Bridge.

Despite variations in retrieval strategies, all teams, except Pranav, followed a similar paradigm of fine-tuning models with the available training data in a contrastive learning approach. The main difference in their approaches was the way negative samples were selected, with some incorporating NEI-labelled abstracts, while others using the

[12] https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli

[13] https://huggingface.co/naver/splade-v3

[14] https://deepmind.google/models/gemini/flash/



(a) Subtask I Scores

(b) Subtask II Scores

Figure 3: Results of participants who submitted system descriptions compared to our baselines for subtask I (left) and subtask II (right). Subtask I scores are reported using Recall@K ($K = 2, 5, 10$), Bpref, and the average-based SubtaskI-Score. Subtask II scores are reported using weighted metrics of Precision, Recall, and F1, along with SubtaskII-Score which is the sum of Recall@10 from subtask I and F1 from subtask II.

least relevant abstracts from their own retrieval approach. We hypothesise that this enabled models to distinguish subtle semantic differences in scientific discourse (Zhan et al., 2021). Notably, the top two systems fine-tuned the cross-encoders, while the others did so on their dense retrieval models.

When comparing systems, we additionally note the impact of retrieval depth and recall preservation. The top-ranked system retrieved up to 5000 abstracts per claim before reranking, enabling a high coverage of potentially relevant documents. In contrast, systems that retrieved a limited number of abstracts early on could have missed documents, impacting the effectiveness of reranking. This highlights that in tasks where relevant evidence is sparse and semantically complex, such as scientific abstracts, high recall in retrieval is effective.

For subtask II, all leaderboard results employed an LLM classification approach, resulting in relatively small margins in their scores. Notably, the top two teams used closed-source, commercial LLMs, while the third ranked team and the baseline employed open-source models. That being said, team EFC showed that a more lightweight architecture, fine-tuned correctly, can still yield competitive results, highlighted by the results they achieved using DeBERTa. This emphasises the practical trade-off between performance and efficiency, which is an important consideration for real-world applications such as content moderation or misinformation detection. In such scenarios, latency, scalability, and interpretability matter. Thus, systems optimised for low-resource settings remain very relevant, while other systems that employ commercial LLMs might be less useful.

9 Conclusion

This paper presented the ClimateCheck shared task, which focused on fact-checking claims from social media about climate change against scholarly articles. The task ran during April/May 2025 and was hosted as part of the 5th SDP Workshop in 2025. Given a claim, two subtasks were available: (I) Retrieving the top 10 most relevant (i. e., evidentiary) abstracts, and (II) Classifying the veracity of the claim given the abstract. The first subtask was evaluated using Recall@K ($K = 2, 5, 10$) and Bpref, while the second using F1 with additional scaling based on correctly retrieved abstracts. The task received ten leaderboard submissions, three of which for both subtasks. Participants explored a wide range of retrieval and classification strategies, including sparse and dense retrieval fusion, supervised reranking with cross-encoders, prompt-based classification with LLMs, and fine-tuned transformer classifiers. Despite methodological differences, the most effective systems shared an emphasis on high-recall retrieval, robust reranking, and careful label calibration. The ClimateCheck datasets are publicly available,[15,16] and a test suite can be accessed for further submissions by the community.[17] While the task results are encouraging, it remains an open question whether these systems are reliable enough for practical deployment. Key open challenges include ensuring system robustness under noisy or multilingual in-

[15] https://huggingface.co/datasets/rabuahmad/climatecheck

[16] https://huggingface.co/datasets/rabuahmad/climatecheck_publications_corpus

[17] https://www.codabench.org/competitions/8304/

put, reducing inference latency for real-time use, and scaling evidence retrieval across large scholarly corpora. Addressing these challenges will be essential to transition from prototype systems to real-world fact-checking tools that can support climate literacy and policy discourse.

Limitations

Although the ClimateCheck task provides a valuable benchmark for evaluating retrieval-augmented fact-checking systems in the climate science domain, several limitations should be noted. First, the evaluation was conducted at the abstract level, which may not fully capture the granularity needed for real-world scientific fact-checking, where evidence often resides at the sentence or paragraph level. This limited both the precision of retrieval and the interpretability of classification outputs.

Moreover, although the task focused on social media claims, the claims were presented in isolation, without access to contextual metadata (such as source, post history, or surrounding discourse). As a result, systems could not leverage pragmatic or contextual cues that are often important in assessing claim intent or credibility in practice.

While the task encouraged participation in both subtasks, only a small subset of teams did so, limiting the ability to assess full-pipeline performance across systems. Additionally, some systems relied on commercial LLMs, which, while effective, reduce reproducibility and raise concerns around fairness in evaluation due to their proprietary nature and limited accessibility.

The annotated training data is relatively limited in size and scope, covering a restricted set of claims and evidence pairs. Although sufficient to train and evaluate retrieval and classification models, further scaling is needed to support generalisation across claim types and evidence complexity. More training data is planned to be annotated in the next months and released as an updated version of the ClimateCheck dataset.

Finally, a notable limitation in the evaluation setup stems from the iterative annotation process, which introduced an inherent bias toward teams that submitted results early and consistently. Throughout the competition, additional evidence annotations were guided by intermediate system outputs, meaning that teams whose systems were included in early and repeated annotation rounds had the advantage of gold testing data that better reflected their own retrieval outputs. Unsurprisingly, the top four teams participated from the beginning and were included in nearly all annotation iterations. In contrast, team Pranav stands out as the only team to outperform the baseline without ever being included in the additional annotation cycles. This highlights how annotation strategies can unintentionally reinforce system-specific retrieval patterns, favouring early participants and potentially underestimating the performance of latecomers.

Ethical Statement

Our annotators were compensated through a typical payment scheme and have been informed about the further use of their annotations. The claims used in the task do not contain sensitive or personal information and are collected from open-source datasets. Due to preprocessing, real claims from social media cannot be traced back to their original posts. We additionally emphasise that automated fact-checking systems are not a substitute for expert judgement and should be deployed with appropriate human oversight.

Acknowledgements

This work was supported by the consortium NFDI for Data Science and Artificial Intelligence (NFDI4DS)[18] 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, and Shuyue Qu for their work.

References

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 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria.

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Ahmed Al-Rawi, Derrick O'Keefe, Oumar Kane, and Aimé-Jules Bizimana. 2021. Twitter's fake news dis-

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

courses around climate change and global warming. *Frontiers in Communication*, 6:729818.

Firoj Alam, Julia Maria Struß, Tanmoy Chakraborty, Stefan Dietze, Salim Hafid, Katerina Korre, Arianna Muti, Preslav Nakov, Federico Ruggeri, Sebastian Schellhammer, and 1 others. 2025. The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval. In *European Conference on Information Retrieval*, pages 467–478. Springer.

Monther Aldwairi and Ali Alwahedi. 2018. Detecting fake news in social media networks. *Procedia Computer Science*, 141:215–222.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. 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, Hong Kong, China. Association for Computational Linguistics.

Chris Buckley and Ellen M Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1870–1879. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024b. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

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*.

Mahmoud Fathallah, Nagwa El-Makky, and Marwan Torki. 2025. AlexUNLP-FMT at ClimateCheck shared task: Hybrid retrieval with adaptive similarity graph-based reranking for climate-related social media claims fact checking. In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria.

Michele Filannino and Özlem Uzuner. 2018. Advancing the state of the art in clinical natural language processing through shared tasks. *Yearbook of medical informatics*, 27(1):184–192.

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, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Salim Hafid, Sebastian Schellhammer, Sandra Bringay, Konstantin Todorov, and Stefan Dietze. 2022. Scitweets-a dataset and annotation framework for detecting scientific online discourse. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3988–3992.

Anthony James Hughes and Xingyi Song. 2024. Identifying and aligning medical claims made on social media with medical evidence. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8580–8593, Torino, Italia. ELRA and ICCL.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Anna Kiepura and Jessica Lam. 2025. Climate-Check2025: Multi-stage retrieval meets llms for automated scientfic fact-checking. In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria.

Carlos Lassance, Hervé Déjean, Thibault Formal, and Stéphane Clinchant. 2024. SPLADE-v3: New baselines for SPLADE. *arXiv preprint arXiv:2403.06789*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel 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, Online. Association for Computational Linguistics.

Preslav Nakov, Alberto Barrón-Cedeno, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings 9*, pages 372–387. Springer.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6859–6866. AAAI Press.

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)*.

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 and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos. 2024. The automated verification of textual claims (AVeriTeC) shared task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*.

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, Torino, Italia. ELRA and ICCL.

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, Online. Association for Computational Linguistics.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Max Upravitelev, Nicolau Duran-Silva, Christian Woerle, Giuseppe Guarino, Jing Yang Salar Mohtaj, Veronika Solopova, and Vera Schmitt. 2025. Comparing LLMs and BERT-based classifiers for resource-sensitive claim verification in social media. In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria.

David Wadden and Kyle Lo. 2021. Overview and insights from the SCIVER shared task on scientific claim verification. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 124–129, Online. Association for Computational Linguistics.

Junjun Wang, Kunlong Chen, Zhaoqun Chen, Peng He, and Wenlu Zheng. 2025. Winning ClimateCheck: A multi-stage system with BM25, BGE-reranker ensembles, and LLM-based analysis for scientific abstract retrieval. In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, and 1 others. 2024. Yi: Open foundation models by 01.ai. *arXiv preprint arXiv:2403.04652*.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 1503–1512. ACM.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024. Jasper and Stella: distillation of SOTA embedding models. *arXiv preprint arXiv:2412.19048*.

A Dataset Samples

Table 6 presents five random claims extracted from the test set of the ClimateCheck dataset. Each claim is presented with the top five abstracts retrieved by the three teams that participated in the two shared task subtasks: Ant Bridge, akiepura_ljam, and EFC. Each abstract is followed by a symbol indicating whether it supports, refutes, or does not have enough information about the claim, according to the results of the team's subtask II labels.

B Iterative Human Annotation Process

During the testing phase of the competition, additional documents were collected based on submissions to mitigate bias in the gold testing data. We did so using the following timeline:

- **Week 1, submissions until April 25, 2025**: we had two active teams: Ant Bridge and AlexUNLP-FMT, thus filtering based on an agreement threshold of 2 without further filtering based on ranking. We extracted the following runs: 275408 and 272964, resulting in 405 additional annotated documents.
- **Week 2, submissions until May 2, 2025**: we received submissions from 3 active participants: Ant Bridge, AlexUNLP-FMT, and akiepura_jlam. We filtered pairs for annotation with an agreement between at least two teams and a minimum rank of 8 across all teams. This helped us manage the annotation workload while still maintaining a fairer evaluation strategy, taking into account all active teams. The following runs were extracted: 279364, 280185, and 280233. As a result, 351 additional pairs were annotated.
- **Week 3, submissions until May 13, 2025**: seven users were active: Ant Bridge, AlexUNLP-FMT, akiepura_jlam, gmguarino, salarmohtaj, nicolauduran45, and EFC, from which we filtered based on an agreement of at least three systems with no further ranking filtering. The following runs were extracted: 285646, 285887, 286061, 286273, 286663, 286806, 286836. This resulted in 477 new annotated claim-abstract pairs.

Overall, the full process resulted in 1,233 additional human-annotate claim-abstract pairs for the 176 unique claims in the test set.

Input (Claim)	Output (Publications)		
	Ant Bridge	akiepura_jlam	EFC
People make it seem like we can change our energy habits, which is quite difficult.	1. Maréchal, 2014 ✓ 2. Jaccard, 2020 ✓ 3. De Vries et al., 2011 ◦ 4. Jans et al., 2018 ✓ 5. Malott, 2017 ◦	1. Jaccard, 2020 ✗ 2. Maréchal, 2014 ✓ 3. Horgan et al. 2016 ✓ 4. De Vries et al., 2011 ✓ 5. Bloodhart et al., 2013 ✗	1. Maréchal, 2014 ✓ 2. Jaccard, 2020 ✓ 3. Horgan et al. 2016 ✓ 4. Viguié et al., 2020 ✓ 5. Welton, 2018 ◦
Greenhouse gases from our actions are a major factor in warming our planet.	1. Nadeau et al., 2021 ✓ 2. Simkins, 1991 ✓ 3. Feely et al., 2015 ✓ 4. Verma, 2021 ✓ 5. Solomon et al., 2010 ✓	1. Al-Ghussain, 2018 ✓ 2. Simkins, 1991 ✓ 3. Haines & Patz, 2004 ✓ 4. Nadeau et al., 2021 ✓ 5. Miller et al., 2008 ✓	1. Nadeau et al., 2021 ✓ 2. Simkins, 1991 ✓ 3. Gadani & Vyas, 2011 ✓ 4. Haines & Patz, 2004 ✓ 5. Giudice et al., 2021 ✓
Burning biomass is a source of air pollution.	1. Rogers et al., 2020 ✓ 2. Huang et al., 2016 ✓ 3. Naik et al., 2007 ✓ 4. Corsini et al., 2019 ✓ 5. Sigsgaard et al., 2015 ✓	1. Rogers et al., 2020 ✓ 2. Corsini et al., 2019 ✓ 3. Naik et al., 2007 ✓ 4. Sigsgaard et al., 2015 ✓ 5. Unosson et al., 2013 ✓	1. Naik et al., 2007 ✓ 2. Corsini et al., 2019 ✓ 3. Rogers et al., 2020 ✓ 4. Li et al., 2019 ✓ 5. Huang et al., 2016 ✓
heat waves have been on a downward trend both in the US and globally #ClimateChangeFacts	1. Peterson et al., 2013 ✗ 2. Ceccherini et al., 2016 ✗ 3. Bumbaco et al., 2013 ✗ 4. Cao et al., 2021 ◦ 5. Li & Amatus., 2020 ✗	1. Peterson et al., 2013 ✗ 2. Ceccherini et al., 2016 ✗ 3. Bumbaco et al., 2013 ✗ 4. Chase et al., 2006 ✗ 5. Mo & Lettenmaier, 2015 ◦	1.Peterson et al., 2013 ✗ 2. Huang et al., 2021 ✗ 3. Ceccherini et al., 2016 ✗ 4. Bumbaco et al., 2013 ✗ 5. Mo & Lettenmaier, 2015 ◦
Apparently, ice caps are at record levels now, despite predictions of melting.	1. Thompson, 2017 ✗ 2. Anderson et al., 2008 ✗ 3. Isaksson et al., 2005 ◦ 4. NEEM community members, 2013 ✗ 5. Thompson et al., 2021 ✗	1. Devasthale et al., 2013 ✗ 2. Taranczewski et al., 2019 ✗ 3. Graeter et al., 2018 ✗ 4. Thompson, 2017 ✗ 5. Massonnet et al., 2023 ✗	1. Edwards et al., 2019 ◦ 2. Taranczewski et al., 2019 ✗ 3. Hanna et al., 2013 ◦ 4. Devasthale et al., 2013 ✗ 5. Graeter et al., 2018 ✗

Table 6: Sample of five random claims from the test set along with the top five retrieved abstracts from each one of the three teams that participated in both subtasks. Each abstract is followed by a symbol denoting the annotation label given to the claim-abstract pair: ✓ = Supports, ✗ = Refutes, ◦= NEI.