

ClimateCheck 2026: Scientific Fact-Checking and Disinformation Narrative Classification of Climate-related Claims

Raia Abu Ahmad^{1,2} Max Upravitelev^{1,2} Aida Usmanova³
Veronika Solopova^{1,2} Georg Rehm^{1,4}

¹Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Germany

²Technische Universität Berlin, Germany ³Leuphana Universität Lüneburg, Germany

⁴Humboldt-Universität zu Berlin, Germany

Corresponding author: raia.abu_ahmad@dfki.de

Abstract

Automatically verifying climate-related claims against scientific literature is a challenging task, complicated by the specialised nature of scholarly evidence and the diversity of rhetorical strategies underlying climate disinformation. ClimateCheck 2026 is the second iteration of a shared task addressing this challenge, expanding on the 2025 edition with tripled training data and a new disinformation narrative classification task. Running from January to February 2026 on the CodaBench platform, the competition attracted 20 registered participants and 8 leaderboard submissions, with systems combining dense retrieval pipelines, cross-encoder ensembles, and large language models with structured hierarchical reasoning. In addition to standard evaluation metrics (Recall@K and Binary Preference), we adapt an automated framework to assess retrieval quality under incomplete annotations, exposing systematic biases in how conventional metrics rank systems. A cross-task analysis further reveals that not all climate disinformation is equally verifiable, potentially implicating how future fact-checking systems should be designed.

Keywords: scientific fact-checking, disinformation narrative classification, climate change, shared task

1. Introduction

Public understanding of climate change is increasingly shaped by online discourse, the scientific validity of which is often difficult to assess at scale (Fownes et al., 2018). While automatic fact-checking (AFC) has made substantial progress in recent years (Guo et al., 2022), most existing approaches verify claims against general sources such as Wikipedia or web search (Thorne et al., 2018b; Aly et al., 2021; Schlichtkrull et al., 2024). However, recent studies emphasise the importance of using trustworthy evidence sources (Schlichtkrull et al., 2023), which, for scientific topics, means direct engagement with scholarly literature.

Although scientific publications constitute one of the most authoritative forms of evidence for climate-related claims, they remain challenging for AFC systems. This stems from several difficulties when dealing with scholarly document processing, such as the use of in-domain terminology, document length, complex reasoning requirements, connection to other documents through citations, and temporal changes of evidence veracity (Vladika and Matthes, 2023; Deng et al., 2025).

To tackle this, we introduced the ClimateCheck shared task (Abu Ahmad et al., 2025b), focusing on the verification of climate-related social media claims using scientific abstracts as evidence. The task consists of retrieving relevant abstracts for a given claim, and classifying whether the text *supports*, *refutes*, or does *not have enough informa-*

tion (NEI) about the claim, treating the problem at the level of individual claim–abstract pairs (CAPs) instead of producing one final verdict. The competition demonstrated the feasibility of grounding AFC in scholarly knowledge, but also exposed several limitations, including restricted training data, evaluation challenges caused by incomplete annotations, and a growing reliance on computationally expensive large language models (LLMs).

This paper presents the 2026 iteration of ClimateCheck, in which we address these challenges and substantially expand the scope of our dataset. We release approx. three times more training data and introduce a new task, *Disinformation Narrative Classification*, which moves beyond claim-level verification to capture the broader rhetorical structures underlying climate disinformation.

Crucially, all three tasks: retrieval, verification, and disinformation narrative classification, are annotated over the same set of claims, providing a unified dataset for evidence-grounded narrative classification, as demonstrated in Figure 1.¹ To the best of our knowledge, this is the first work that enables systematic investigation of how rhetorical narrative structure interacts with evidence retrieval and veracity classification. Our contributions are threefold:

1. A unified benchmark connecting scholarly evi-

¹Our dataset is publicly available: <https://huggingface.co/datasets/rabuahmad/climatecheck>

dence retrieval, claim verification, and narrative classification.

2. An adapted evaluation framework for incomplete annotations in multi-evidence scientific fact-checking, based on the work proposed by Akhtar et al. (2024).
3. An empirical analysis of system behaviour across retrieval, verification, and narrative tasks, highlighting systematic verification biases and persistent challenges in fine-grained narrative classification.

ClimateCheck 2026 ran from January 15 until February 18 on the CodaBench platform (Xu et al., 2022), with registration from 20 participants and 8 leaderboard submissions across all tasks. Four teams submitted system descriptions, three of which outperformed our baselines. In this paper, we report our dataset development process (§4), evaluation framework (§5), baselines design (§6), participant systems (§7), and error analyses across tasks (§8), which we discuss and conclude in §9 and §10, respectively.

2. Related Work

Shared tasks have played a central role in advancing AFC research by establishing common benchmarks and evaluation frameworks. Early efforts such as FEVER (Thorne et al., 2018b) and FEVEROUS (Aly et al., 2021) focused on verifying claims against Wikipedia, while subsequent work expanded into more specialised retrieval. SciFact (Wadden et al., 2020, 2022), for example, introduced expert-written scientific claims paired with biomedical abstracts, paving the way for the SCIVER shared task (Wadden and Lo, 2021), which demonstrated the particular challenges of reasoning over scholarly publications. More recently, AVeriTeC (Schlichtkrull et al., 2024; Akhtar et al., 2025) shifted the focus to real-world claims verified using open-web evidence, highlighting the importance of trustworthy and traceable sources.

As these tasks have grown in complexity, LLM-based systems have come to dominate leaderboards (Akhtar et al., 2025; Schlichtkrull et al., 2024; Abu Ahmad et al., 2025b). However, recent work suggests that this trend does not consistently yield superior performance, with smaller or task-specialised models remaining competitive in climate-related settings (Calamai et al., 2025; Upravitelev et al., 2025).

Claim-level verification, however, addresses only one dimension of the problem. Climate disinformation rarely operates through isolated false claims, as it tends to cluster around recurring narrative patterns that span many individual statements. Several datasets have been developed to capture these

patterns, grouping climate denial claims by their underlying narrative messages (Coan et al., 2021; Rowlands et al., 2024; Nikolaidis et al., 2025). However, these datasets annotate narrative labels at the claim level without linking them to scientific evidence. To the best of our knowledge, our dataset is the first to connect narrative classification and evidence-based claim verification, enabling the exploration of strategies jointly targeting these tasks.

3. Task Definitions

The 2026 iteration of ClimateCheck consisted of two tasks:

- **Task 1: Abstract retrieval and claim verification:** given a social media claim about climate change, retrieve the top 5 most relevant abstracts from the given publications corpus (task 1.1) and classify each claim-abstract pair as *supports*, *refutes*, or *NEI* (task 1.2).
- **Task 2: Disinformation narrative classification:** given a social media claim about climate change, predict which disinformation narrative(s) are present using the schema proposed by Coan et al. (2021).

Both tasks were evaluated independently, and participants could choose to submit to either or both. Participants were allowed to use external datasets and apply augmentation methods.

4. Dataset Development

Both tasks build on the ClimateCheck dataset introduced in the 2025 iteration (Abu Ahmad et al., 2025a). For task 1, we extend the training split by adding 523 unique English claims, drawn from the original claims pool, and 1897 CAPs, where each claim is linked to 1-5 abstracts. The testing split remains unchanged. The same five annotators who worked on the original dataset also worked on the extension, using the same guidelines described in our previous work. Overall, the dataset for task 1 includes 958 unique claims and 4927 CAPs, each annotated as *supports*, *refutes*, or *NEI*. Table 1 presents the distribution of labels across the training and test splits. The overall inter-annotator agreement (IAA) of the training split including the extension is Cohen’s $\kappa = 0.73$, indicating substantial agreement (Landis and Koch, 1977).

The same 958 unique claims were annotated for task 2, using disinformation narrative labels from the narrative taxonomy and the code book introduced by Coan et al. (2021). The taxonomy distinguishes 27 climate change denial narrative

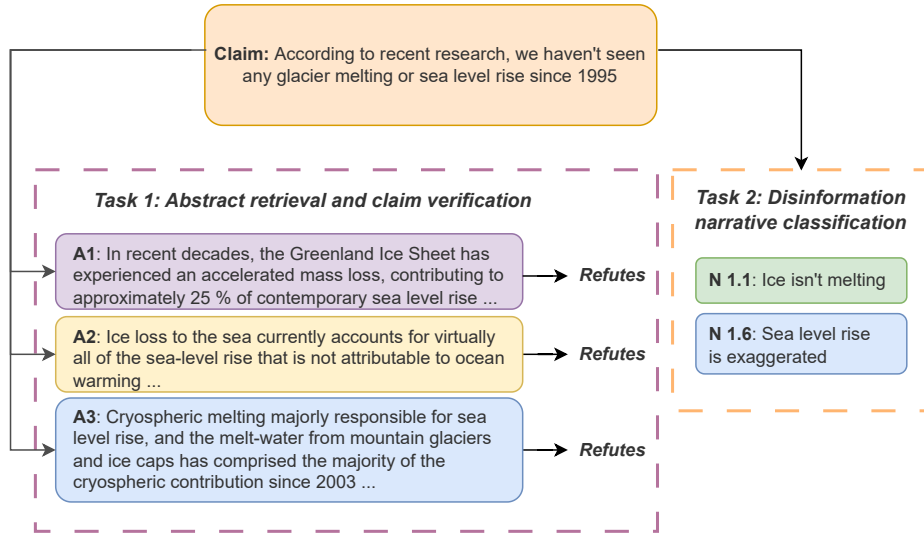


Figure 1: Instance from ClimateCheck 2026. Task 1: Given a claim, systems must retrieve relevant abstracts and use them for verification. Task 2: Given a claim, systems must predict all disinformation narratives associated with it.

Label	Train	Test
Supports	1399	749
Refutes	451	266
Not Enough Information (NEI)	1173	889
Total	3023	1904

Table 1: Label distribution for Task 1 across training and test splits.

labels across 5 main narrative groups, plus a no-disinformation label (33 labels in total). Four annotators applied a multi-label scheme to each claim,² which we adopted since narratives often overlap, and prior work has shown that, in such cases, single-label annotations can lead to errors and downstream classification failures (Calamai et al., 2025). Gold labels were determined by majority vote ($\geq 3/4$ agreement), which was achieved for 83.2% of items (66.6% unanimous). The remaining 16.8% were adjudicated by two of the authors (who were not part of the initial annotator pool) through review and discussion.

We report IAA at three levels of granularity, reflecting the hierarchical nature of the annotation scheme. Treating each (item \times label) pair as an independent binary decision yields an overall Krippendorff’s $\alpha = 0.785$. However, this measure is dominated by the frequent no-disinformation label (71.5% prevalence). The prevalence-weighted average of per-label α values, which better reflects

²Annotation guidelines available at https://github.com/ryabhd/climatecheck/blob/master/narrative_anno_guide.pdf

agreement on the disinformation narratives themselves, is $\alpha = 0.651$, falling just below the $\alpha \geq 0.667$ threshold for content analysis (Krippendorff, 2018). At the level of the five main narrative groups, this value rises to $\alpha = 0.697$, exceeding the threshold and indicating that annotators identify narratives reliably at the top-level, while finer-grained distinctions prove more challenging. This aligns with studies such as Fraile-Hernández et al. (2025), which discuss specific challenges to annotating narrative-related tasks, such as having to account for subjectivity in an interpretive setting. We discuss IAA in further detail and present all labels of the taxonomy in Appendix A.

5. Evaluation Process

5.1. Official Ranking Metrics

Task 1.1: Abstract Retrieval. We evaluate retrieval using Recall@K ($K = 2, 5$) and Binary Preference (Bpref). An abstract is considered evidentiary if it is annotated as *supports* or *refutes* in the gold data.³ The final ranking score for task 1.1 is:

$$\text{Score}_{1.1} = \frac{1}{2} (\text{Recall}@5 + \text{Bpref}).$$

Task 1.2: Claim Verification. For manually annotated CAPs, we compute weighted precision, recall, and F1. Only annotated pairs are considered;

³More detailed definitions of the metrics are available in Appendix B.

unjudged abstracts are ignored. To ensure that verification performance reflects successful evidence retrieval, the final ranking score is defined as:

$$\text{Score}_{1.2} = \text{F1} + \text{Recall@5}.$$

We add R@5 from task 1.1 to penalise systems that achieve a high verification score on a small number of retrieved pairs.

Task 2: Disinformation Narrative Classification. Predictions are evaluated against gold claim labels using macro precision, macro recall, and F1 (macro, micro, and weighted). Due to class imbalance, final rankings are determined by macro F1.

5.2. Evaluation under Incomplete Annotations

Participant systems may retrieve relevant abstracts that are not annotated in the gold data. It is impossible to annotate every CAP, but we still want to reduce bias towards the retrieval approach we based the gold data on. Thus, to evaluate unannotated CAPs, we adapt the Ev²R framework (Akhtar et al., 2024), which has proven to have good correlation with human judgments. Due to limited computational resources, Ev²R scores are computed using the top submission from each team and are reported separately from the official leaderboard.⁴

The Ev²R score combines two complementary components: a reference-based component and a proxy-reference component. The first decomposes both retrieved and reference evidence into atomic facts and evaluates their alignment via precision and recall, measuring how accurately and completely the retrieved evidence reflects the reference. The second uses a fine-tuned DeBERTa model (He et al., 2021) to predict the veracity label for a given claim–evidence pair, with the model’s confidence in the gold label serving as a proxy signal for evidence quality. The final score is a weighted combination of the reference-based F1 score and the proxy-reference confidence score.

Adapted Ev²R Score. The original score assumes a single gold reference per claim, which does not hold in our setting since claims may be linked to multiple gold abstracts with evidence-dependent labels. For a claim c with gold abstracts G_c and a retrieved abstract that is unannotated in the gold data r , we iterate over G_c and retain the gold abstract with the highest alignment to r using the reference-based component:

$$S_{\text{ref}}(r) = \max_{g \in G_c} \text{F1}(r, g).$$

⁴We make our implementation of the framework public for future use: https://github.com/ryabhmd/climatecheck/tree/master/automatic_eval

Let g^* denote this best-aligned abstract and y^* its label. We then compute the proxy-reference component:

$$S_{\text{proxy}}(r) = P_{\theta}(y^* | c, r),$$

measuring how strongly r supports the gold label of the best-aligned abstract. The final adapted Ev²R score per r is:

$$\text{Ev}^2\text{R}(r) = \frac{1}{2}(S_{\text{ref}}(r) + S_{\text{proxy}}(r)).$$

For each claim, we take the maximum Ev²R(r) across all retrieved abstracts, and the submission-level score is the mean over all claims with at least one non-NEI gold abstract (163 of 172 claims in the test set). Claims linked exclusively to NEI abstracts are excluded from this evaluation as they provide no evidential reference. We discuss our implementation in more depth in Appendix C.

Automatic Verification of Unannotated Pairs. For retrieved CAPs outside the gold annotations, we additionally evaluate the predicted label \hat{y} based on the proxy component of the Ev²R framework. We compute a confidence score using the same DeBERTa model:

$$S_{\text{conf}} = P_{\theta}(\hat{y} | c, r),$$

and a consistency score against the label of the best-aligned gold abstract:

$$S_{\text{cons}} = 1[\hat{y} = y^*].$$

The automatic verification score is then:

$$S_{\text{ver}} = \frac{1}{2}(S_{\text{conf}} + S_{\text{cons}})$$

providing an automatic estimate of label plausibility and consistency for unannotated CAPs, allowing us to assess label quality for retrieved abstracts beyond the annotated gold set.

6. Baselines

Task 1: Abstract Retrieval and Claim Verification. The baseline for task 1 is based on the EFC submission from the 2025 iteration (Upravitelev et al., 2025), chosen due to its strong performance relative to its computational requirements. The retrieval component uses a multi-stage pipeline: Abstracts are first retrieved using BM25 (Robertson and Zaragoza, 2009). The top- $k = 1500$ are then re-ranked using semantic similarity computed with a fine-tuned E5 model⁵, which was trained for 3

⁵<https://huggingface.co/intfloat/e5-large-v2>

epochs on our dataset using contrastive learning. Finally, a MiniLM-based cross-encoder⁶ is applied for re-ranking the top- $k = 150$ results, taking the top 5 abstracts per claim as the final predictions. For task 1.2, we use the DeBERTa sequence classification model, fine-tuned on six Natural Language Inference (NLI) datasets: MultiNLI (Williams et al., 2018), ANLI (Nie et al., 2020), LingNLI (Parrish et al., 2021), WANLI (Liu et al., 2022), FEVER NLI (Nie et al., 2019), and the ClimateCheck training split. We optimise for highest minimum accuracy per label to mitigate class imbalance.

Task 2: Disinformation Narrative Classification. The baseline employs Qwen3-8B (Yang et al., 2025) fine-tuned in a supervised instruction-following setup. Qwen3 is chosen due to good performance on a public leaderboard,⁷ specifically targeting the evaluation of smaller and more efficient LLMs. The training data is constructed by converting claim–narrative pairs into chat-style instruction–response examples using the CARDS taxonomy (see prompt in Appendix D). Fine-tuning is performed using parameter-efficient Low-Rank Adaptation (LoRA, Hu et al., 2022) with standard cross-entropy loss on the assistant responses. The model predicts one or more taxonomy codes per claim in a constrained generation format.

Our code for implementing baseline models for both tasks is publicly available to enable reproducibility.⁸

7. Submitted Systems and Results

We received 8 leaderboard submissions across both tasks, the results of which are shown in Table 2 for task 1 and Table 3 for task 2. We also report scores using the Ev²R framework for task 1 submissions in Table 4. In what follows, we briefly describe the system descriptions we received from four teams.

ClimateSense (Ehrhart et al., 2026). For task 1, the team employed a three-stage pipeline, following the winning team of the 2025 iteration (Wang et al., 2025). Initial candidate retrieval used BM25, followed by an ensemble of five fine-tuned BGE re-rankers (Chen et al., 2024) trained with hard negative mining, and aggregated via Reciprocal Rank Fusion. For claim verification, the team opted for

⁶<https://huggingface.co/cross-encoder/ms-marco-MiniLM-L12-v2>

⁷<https://huggingface.co/spaces/k-mktr/gpu-poor-llm-arena>

⁸<https://github.com/XplaiNLP/ClimateCheck-2026-Baseline/>

zero-shot (ZS) classification using gpt-oss-120b⁹. A fine-tuned DeBERTa-based cross-encoder was also explored but underperformed compared to the ZS LLM approach. For task 2, they adopted a ZS approach using GPT 5.1¹⁰. The prompt incorporated the full CARDS taxonomy definitions, negative criteria for each category, specificity guidelines, and five-step chain-of-thought (CoT) reasoning.

DFKI-IML (Liang et al., 2026). The team focused mainly on task 1.2, adopting a similar retrieval pipeline as the baseline and directing their attention to claim verification. They compared two self-explainable inference paradigms: Intermediate Reasoning (IR), where the model generates an explanation before predicting the entailment label, and Post-hoc Rationalization (PR), where the label is predicted before the explanation. Three models were evaluated under both paradigms in a ZS setting: GPT-4o-mini (Achiam et al., 2023), Phi-3-Medium (Abdin et al., 2024), and Mistral-Nemo¹¹. IR yielded more stable and consistent results across all three models, with GPT-4o-mini achieving the highest scores.

ahilbert (Hilbert et al., 2026). The team focused exclusively on task 2, using Qwen3-8B as a fixed backbone to compare approaches. They investigated three directions: data augmentation, prompt engineering, and reinforcement learning via Group-Relative Policy Optimization (GRPO, Shao et al., 2024). They augmented the data with synthetic claims, applying skewed inverse-frequency reweighting to oversample minority narratives. For prompt engineering, they compared three strategies: a simple direct instruction prompt, a hierarchical two-turn approach, and a CoT prompt that encourages the same hierarchical reasoning within a single forward pass through claim decomposition, high-level group selection, and sub-label assignment. The best-performing setup was ZS CoT prompting combined with Qwen3’s native reasoning mode. The team also reported inference-time emissions, showing a clear trade-off between reasoning-based performance and computational cost, with reasoning-heavy approaches consuming roughly 25 times more energy than ZS inference.

XplaiNLP (Foroutan et al., 2026). The team worked on task 2, investigating both encoder-based and decoder-only approaches. For the former, they fine-tuned ModernBERT-large (Warner et al., 2025),

⁹<https://huggingface.co/openai/gpt-oss-120b>

¹⁰<https://openai.com/index/gpt-5-1/>

¹¹<https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407>

Team	Task 1.1				Task 1.2			
	R@2	R@5	Bpref	Score _{1.1}	P	R	F1	Score _{1.2}
Baseline	0.213	0.403	0.459	0.431	0.683	0.682	0.679	1.082
Ant Bridge (Wang et al., 2025) [†]	0.218	0.451	0.495	0.473	0.729	0.726	0.725	1.176
ClimateSense (Ehrhart et al., 2026)	0.221	0.443	0.489	0.466	0.742	0.744	0.740	1.183
berkbubus	0.206	0.429	0.453	0.441	0.391	0.414	0.328	0.757
DFKI-IML (Liang et al., 2026)	0.193	0.352	0.401	0.377	0.670	0.641	0.620	0.972
gardlz	0.193	0.364	0.314	0.340	0.581	0.593	0.579	0.943
ytsoneva	0.060	0.094	0.116	0.105	0.269	0.176	0.148	0.242

[†] Winning team of the 2025 iteration on the same test set; included for reference only.

Table 2: Results for Task 1. Task 1.1 is evaluated using Recall@2 (R@2), Recall@5 (R@5), Binary Preference (Bpref), and Score_{1.1}. Task 1.2 is evaluated using weighted Precision (P), Recall (R), F1, and Score_{1.2}. Best results among 2026 submissions are shown in **bold**. Shaded cells indicate scores that outperform the baseline.

Team	P (macro)	R (macro)	F1 (macro)	F1 (micro)	F1 (weighted)
Baseline	0.530	0.574	0.514	0.798	0.784
ahilbert (Hilbert et al., 2026)	0.707	0.631	0.625	0.844	0.821
XplaiNLP (Foroutan et al., 2026)	0.625	0.640	0.597	0.801	0.806
ClimateSense (Ehrhart et al., 2026)	0.670	0.568	0.583	0.876	0.844
alexsiakalou	0.574	0.586	0.548	0.844	0.837

Table 3: Results for Task 2. Systems are evaluated using macro-averaged Precision (P), Recall (R), and F1, as well as micro- and weighted F1. Macro-F1 is the official ranking metric, reflecting performance across all narrative categories regardless of class frequency. Best results are shown in **bold**. Shaded cells indicate scores that outperform the baseline.

RoBERTa (Liu et al., 2019), and DistilBERT (Sanh et al., 2019) on the the training data, additionally using the Augmented CARDS dataset (Rojas et al., 2024) to mitigate class imbalance. ModernBERT with the augmented dataset achieved the strongest results, surpassing our baseline at a fraction of the computational cost. For decoder-only models, the team fine-tuned Qwen3-8B under three instruction tuning strategies: prompt enhancement with in-context examples, hierarchical instruction tuning where top-level and fine-grained categories were predicted in two sequential stages, and retrieval-augmented instruction tuning where a cross-encoder first retrieved the ten most semantically similar narrative descriptions from the CARDS taxonomy. The retrieval-augmented approach consistently outperformed the other strategies.

8. Error Analysis

Retrieval Difficulty. Figure 2 shows the distribution of per-claim average Recall@5 across all submitted systems, sorted in ascending order. The distribution is notably gradual, suggesting that retrieval difficulty is not a binary easy/hard problem. Only one claim achieves Recall@5=0 across all

systems,¹² despite being linked to evidentiary abstract in the gold data. A possible reason could be a lexical mismatch between the claim and the available abstracts, preventing sparse retrieval systems (i.e. BM25, used by all teams) from surfacing relevant evidence. The vast majority of claims (n=137) fall in the mid-range between 0 and 0.5, and the easiest claims (n=25, R@5_{≥0.5}) are rarely solved perfectly by all systems, indicating that no system consistently achieves high recall, even on claims where retrieval is easiest.

Verification Label Confusion. Figure 3 shows normalised confusion matrices for submitted systems from our baseline, ClimateSense, and DFKI-IML on task 1.2.¹³ A clear pattern emerges across systems: *Refutes* is the hardest label to predict correctly, with diagonal values ranging from 0.05 to 0.68. This is consistent with the broader NLI literature, showing that refutation requires stronger evidence-grounded reasoning than support or abstention (Atanasova et al., 2022; Thorne et al., 2018a). The most common error is the confusion

¹²“NOAA has adjusted past temperatures to look colder and recent ones warmer”

¹³We show further systems, whose teams did not submit reports, in Appendix E.

Team	Ev ² R(r)	S _{ver}	# Evaluated Claims	Avg. Unannotated Abstracts per Claim
Baseline	0.606	0.600	112	1.92
ClimateSense (Ehrhart et al., 2026)	0.559	0.559	111	1.77
berkbubus	0.554	0.468	106	1.67
DFKI-IML (Liang et al., 2026)	0.568	0.527	135	1.97
gardlz	0.585	0.528	134	2.16
ytsoeva	0.672	0.143	163	4.06

Table 4: Ev²R evaluation results for unannotated abstracts retrieved in Task 1. Ev²R(r) is the mean of the reference-based and proxy-reference components evaluating task 1.1. S_{ver} is the added verification score evaluating task 1.2. Only the final leaderboard submission per team is evaluated.

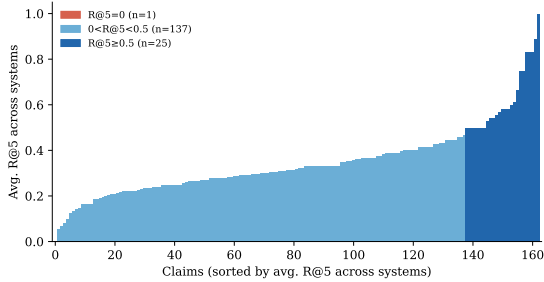


Figure 2: Per-claim average Recall@5 across all submitted systems, sorted in ascending order. Colours indicate retrieval difficulty.

of *NEI* with *Supports*: the baseline misclassifies 29% of such instances, ClimateSense 27%, and DFKI-IML 48%. This suggests that systems tend to interpret partial or weakly relevant evidence as supporting rather than insufficient. We also note a particular error pattern in DFKI-IML, where 39% of gold *Refutes* instances are predicted as *Supports*, meaning that the system flips its interpretation of refuting evidence, potentially leading to false affirmation of misinformation.

Narrative Classification. We analyse error patterns of the submitted systems on Task 2. Figure 4 shows per-label F1 scores for each system alongside IAA (Krippendorff’s α). On average, 80.9% of test claims received an exact-match prediction, 3.5% were *partially correct* (correct top-level category, wrong sub-narrative), and 15.6% were assigned a completely wrong top-level category. Of wrong predictions, 81% crossed top-level category boundaries while only 19% stayed within the correct category, indicating that identifying the broad narrative type is the primary challenge rather than distinguishing sub-narratives. Systems also showed a consistent under-prediction tendency: 4.8% of predictions contained too few labels versus only 0.6% with too many, with 0_0 (*no disinformation*) being the most common false prediction. Labels that were difficult for annotators were also difficult

for systems. Among labels with sufficient test support, Krippendorff’s α and mean system F1 show a perfect rank correlation (Spearman $\rho=1.0$, $n=4$), though the small number of qualifying labels limits statistical power. This pattern is visible in Figure 4: label 5_1 ($\alpha=0.55$) has both the lowest and most variable system F1 (0.22–0.62), while 0_0 ($\alpha=0.72$) is consistently well-handled ($F_1 \geq 0.89$). Label 2_1 (*natural cycles*; $\alpha=0.56$) shows the widest inter-system spread (0.46–0.92), suggesting that system architecture matters most for labels with moderate annotator agreement. Seven claims (4.1%) were misclassified by all five systems ($F_1=0.0$), and a majority-vote ensemble scored below the best individual system (84.3% vs. 85.5% exact match), indicating correlated errors.

Cross-task Analysis. To investigate whether disinformation narratives affects retrieval and verification difficulty, we analyse system performance across the five main narrative groups using the two systems that submitted predictions for both tasks: our baseline and ClimateSense. We first note that retrieval is narrative-agnostic, with Recall@5 results being consistently high across all groups.¹⁴ This indicates that relevant scientific evidence is retrievable regardless of the type of disinformation narrative a claim embodies. However, we see a different pattern when looking at verification difficulty, signaling that it is strongly narrative-dependent. Looking at the most difficult label to verify, *Refutes*, we note a clear gradient where claims belonging to Group 3 (*climate impacts are not bad*) are the easiest to refute, while Groups 4 (*climate solutions won’t work*) and 5 (*climate movement/science is unreliable*) are systematically the hardest (See Figure 5). This result is structurally motivated, since Group 4 claims are primarily normative or economic in nature, and Group 5 claims are epistemological, attacking the credibility of science itself rather than making falsifiable empirical assertions. Thus, using a scientific abstract to refute a claim that science is

¹⁴See Appendix E for full results.

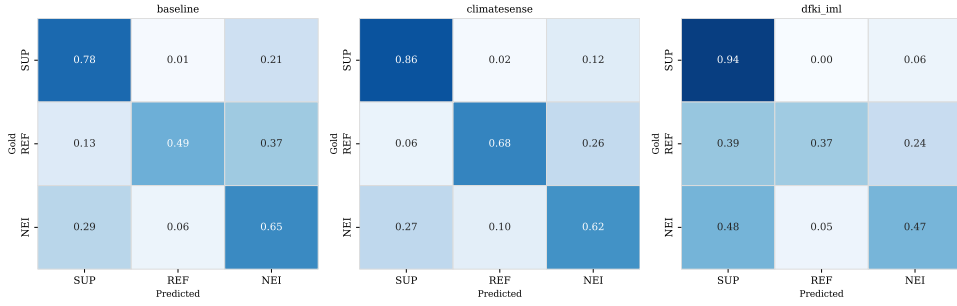


Figure 3: Confusion matrices for the baseline, ClimateSense, and DFKI-IML predictions on task 1.2, normalised by claim. SUP = Supports, REF = Refutes, NEI = Not Enough Information.

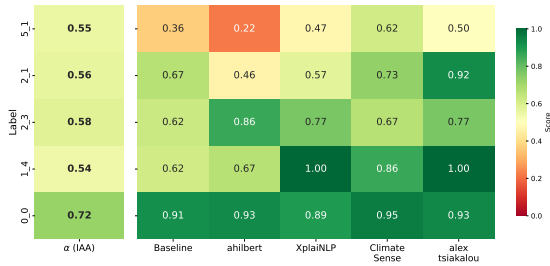


Figure 4: Per-label F1 scores across participating systems and IAA (Krippendorff’s α , leftmost column). Labels sorted by mean system F1 (ascending); only labels with more than 3 instances in the test set are shown.

unreliable is self-referentially problematic: the evidence source is precisely what the claim calls into question. This represents a structural limitation of evidence-grounded AFC that is difficult to resolve through improved retrieval or stronger verification models alone.

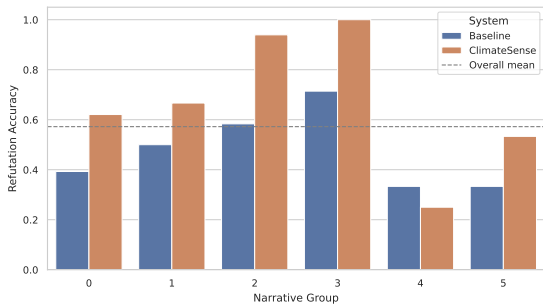


Figure 5: Refutation accuracy by narrative group based on the CARDS taxonomy for the baseline and ClimateSense systems.

9. Discussion

System Trends. Results across both tasks reveal a consistent pattern: structured reasoning and task-specific architectural choices matter more

than model size or data volume alone. For task 1, the strongest retrieval systems rely on multi-stage dense re-ranking pipelines with cross-encoder ensembles, echoing findings from the 2025 iteration. Notably, although ClimateSense adopts a retrieval approach similar to the 2025 winning team, their retrieval performance remains lower despite the availability of more training data in this iteration, suggesting that implementation details and training methodology are more important than data volume at this scale. For task 2, the most effective approaches share a common thread: hierarchical or structured reasoning that identifies coarse-grained narrative groups before committing to fine-grained sub-labels. Team ahilbert’s results further highlight an important practical consideration, raising questions about the deployability of LLM-based verification at scale in real-world AFC pipelines.

Evaluation Beyond Gold Annotations. The Ev^2R results reveal trends that the official leaderboard alone does not capture. Most notably, team *ytsonova*, which ranks last on the official retrieval metric, achieves the highest $Ev^2R(r)$ score. This is largely explained by its retrieval behaviour: it is the only system that retrieves at least one unannotated abstract for all 163 evaluated claims, with an average of more than 4 out of 5 retrieved abstracts per claim falling outside the gold annotations. This means the official metrics penalise this system almost entirely for retrieving evidence that was never judged, rather than for retrieving poor evidence. The Ev^2R score suggests that this evidence is in fact of reasonable quality, a signal that wouldn’t be visible under standard evaluation alone. Beyond this case, the baseline achieves the highest Ev^2R score among the remaining systems despite not ranking first on the official leaderboard, further suggesting that systems optimising directly for gold annotations may be learning retrieval biases introduced by the annotation process. Taken together, these results reinforce the importance of complementing recall-based metrics with automatic evaluation frameworks whenever annotation complete-

ness cannot be guaranteed, which is a particularly pressing concern in AFC, where exhaustive annotation is rarely feasible.

The Cross-Task Connection. Our cross-task analysis reveals that verification difficulty is strongly narrative-dependent, with narrative groups 4 and 5 proving systematically resistant to evidence-based refutation. This finding has a direct implication for system design: a pipeline that first classifies the disinformation narrative of a claim and then conditions its verification strategy on that classification is a natural and promising direction. For claims that make concrete empirical assertions, standard evidence retrieval and NLI-based verification may be sufficient. For other claims, however, the structural mismatch between the claim type and the evidence source suggests that different strategies might be needed, such as specialised evidence types beyond scholarly abstracts, or flagging for human review rather than automated verdict. Crucially, none of the submitted systems attempted to exploit the connection between the two tasks, leaving narrative-aware verification an open and promising direction that can be explored further.

10. Conclusion

We presented ClimateCheck 2026, expanding the shared task with tripled training data, a new disinformation narrative classification task, and a unified benchmark connecting evidence retrieval, claim verification, and narrative classification over a common set of claims. Three of four participating teams outperformed our baselines, with multi-stage dense retrieval pipelines and structured hierarchical reasoning proving the most effective strategies for tasks 1 and 2 respectively. Beyond system-level results, our cross-task analysis reveals a structural limitation of evidence-grounded AFC: while retrieval difficulty is narrative-agnostic, verification difficulty is strongly narrative-dependent, with epistemological and normative claims proving more difficult to scholarly refutation. This motivates narrative-aware verification as a promising direction for future work, which our unified benchmark is uniquely positioned to support.

11. Limitations

The ClimateCheck dataset is restricted to English-language claims only, limiting the applicability of trained systems to the broader multilingual landscape of online climate discourse. In addition, the claim verification task, as formulated here, treats verification as a pairwise problem between a single claim and a single abstract. This is a simplification that enables scalable annotation, but does not

reflect the full complexity of scientific AFC, where a verdict may depend on evidence across multiple documents, conflicting findings, and following chains of citations. Future iterations could explore multi-hop verification settings, where evidence from several abstracts must be jointly considered to reach a final verdict.

In terms of participating systems, we note that, for task 1, teams largely replicated or incrementally extended the methods that proved successful in the 2025 iteration of the shared task, with little methodological novelty in the retrieval stage. Additionally, despite all claims being annotated for both verification and disinformation narrative labels, no team explored whether narrative predictions could inform veracity classification or vice versa. Given that disinformation narratives carry implicit expectations about the direction of evidence, narrative-aware verification represents a promising direction.

Finally, the annotation of fine-grained disinformation narratives remains challenging even for human annotators, as reflected in the below-threshold IAA for several categories. This inherent subjectivity sets an upper bound on what automated systems can be expected to achieve, and suggests that future work may benefit from the introduction of hierarchical evaluation metrics that reward partial credit for correct top-level classification.

12. Acknowledgments

This work was supported by the consortium NFDI for Data Science and Artificial Intelligence (NFDI4DS)¹⁵ 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). Furthermore, the work was partly performed in the scope of the projects “VeraXtract” (reference: 16IS24066) and “news-polygraph” (reference: 03RU2U151C) funded by the German Federal Ministry for Research, Technology and Aeronautics (BMFTR). We thank the annotators: Emmanuella Asante, Farzaneh Hafezi, Senuri Jayawardena, and Shuyue Qu for working on the extension of the task 1 data, and Berk Bubus, Paul-Conrad Feig, Neda Foroutan, Alexandra Tsiakalou for annotating the task 2 data. We also thank Nikolas Rauscher for helping with training the DeBERTa model used for computing the Ev²R scores.

¹⁵<https://www.nfdi4datascience.de>

13. Bibliographical References

- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Pra-neetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#).
- Raia Abu Ahmad, Aida Usmanova, and Georg Rehm. 2025a. [The ClimateCheck dataset: Mapping social media claims about climate change to corresponding scholarly articles](#). In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 42–56, Vienna, Austria. Association for Computational Linguistics.
- Raia Abu Ahmad, Aida Usmanova, and Georg Rehm. 2025b. [The ClimateCheck shared task: Scientific fact-checking of social media claims about climate change](#). In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 263–275, Vienna, Austria. Association for Computational Linguistics.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Floren-cia Leoni Aleman, Diogo Almeida, Janko Al-tenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Mubashara Akhtar, Rami Aly, Yulong Chen, Zhenyun Deng, Michael Schlichtkrull, Chenxi Whitehouse, and Andreas Vlachos. 2025. [The 2nd automated verification of textual claims \(AVeriTeC\) shared task: Open-weights, reproducible and efficient systems](#). In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pages 201–223, Vienna, Austria. Association for Computational Linguistics.
- Mubashara Akhtar, Michael Schlichtkrull, and Andreas Vlachos. 2024. [Ev2r: Evaluating evidence retrieval in automated fact-checking](#). *arXiv preprint arXiv:2411.05375*.
- 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.
- Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. [Fact checking with insufficient evidence](#). *Trans. Assoc. Comput. Linguistics*, 10:746–763.
- 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.
- Tom Calamai, Oana Balalau, and Fabian M. Suchanek. 2025. [Benchmarking the benchmarks: Reproducing climate-related NLP tasks](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17967–18009, Vienna, Austria. Association for Computational Linguistics.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality,](#)

- multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- 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.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *arXiv preprint arXiv:2507.06261*.
- Xingyu Deng, Xi Wang, and Mark Stevenson. 2025. [The next phase of scientific fact-checking: Advanced evidence retrieval from complex structured academic papers](#). In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval, ICTIR 2025, Padua, Italy, 18 July 2025*, pages 436–448. ACM.
- Thibault Ehrhart, Grégoire Burel, and Raphael Troncy. 2026. ClimateSense at ClimateCheck 2026. In *Proceedings of the 3rd International Workshop on Natural Scientific Language Processing (NSLP 2026)*, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Neda Foroutan, Alexandra Tsiakalou, and Vera Schmitt. 2026. Retrieval-augmented LLMs and encoder models for multi-label climate disinformation narrative classification. In *Proceedings of the 3rd International Workshop on Natural Scientific Language Processing (NSLP 2026)*, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Jennifer R Fownes, Chao Yu, and Drew B Margolin. 2018. [Twitter and climate change](#). *Sociology Compass*, 12(6):e12587.
- Jesús M. Fraile-Hernández, Anselmo Peñas, and Pablo Moral. 2025. [Automatic identification of narratives: Evaluation framework, annotation methodology, and dataset creation](#). *IEEE Access*, 13:11734–11753.
- Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. [A survey on automated fact-checking](#). *Transactions of the Association for Computational Linguistics*, 10:178–206.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTa: decoding-enhanced BERT with disentangled attention](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Arthur Hilbert, Nils Feldhus, Jing Yang, and Vera Schmitt. 2026. Comparing hierarchical approaches for fine-grained climate disinformation narrative classification. In *Proceedings of the 3rd International Workshop on Natural Scientific Language Processing (NSLP 2026)*, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. 2022. LoRA: Low-rank adaptation of large language models. *Iclr*, 1(2):3.
- Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. [HoVer: A dataset for many-hop fact extraction and claim verification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.
- Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*, 4th edition. SAGE Publications, Thousand Oaks, CA.
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- Siting Liang, Omar Adjali, and Daniel Sonntag. 2026. Towards efficient self-explainable climate-related claim verification with generative models. In *Proceedings of the 3rd International Workshop on Natural Scientific Language Processing (NSLP 2026)*, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. [WANLI: Worker and AI collaboration for natural language inference dataset creation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized BERT pre-training approach](#). *CoRR*, abs/1907.11692.
- Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. [Combining fact extraction and verification with](#)

- neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. [Adversarial NLI: A new benchmark for natural language understanding](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.
- Nikolaos Nikolaidis, Nicolas Stefanovitch, Purificação Silvano, Dimitar Iliyanov Dimitrov, Roman Yangarber, Nuno Guimarães, Elisa Sartori, Ion Androutsopoulos, Preslav Nakov, Giovanni Da San Martino, and Jakub Piskorski. 2025. [Polynarrative: A multilingual, multilabel, multi-domain dataset for narrative extraction from news articles](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31323–31345, Vienna, Austria. Association for Computational Linguistics.
- Alicia Parrish, William Huang, Omar Agha, Soohwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. [Does putting a linguist in the loop improve NLU data collection?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886–4901, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: Bm25 and beyond](#). *Foundations and Trends® in Information Retrieval*, 3(4):333–389.
- Cristian Rojas, Frank Algra-Maschio, Mark Andrejevic, Travis Coan, John Cook, and Yuanfang Li. 2024. [Augmented CARDS: A machine learning approach to identifying triggers of climate change misinformation on twitter](#). *CoRR*, abs/2404.15673.
- Harri Rowlands, Gaku Morio, Dylan Tanner, and Christopher Manning. 2024. [Predicting narratives of climate obstruction in social media advertising](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5547–5558, Bangkok, Thailand. Association for Computational Linguistics.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. [DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter](#). *CoRR*, abs/1910.01108.
- 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)*, pages 1–26, Miami, Florida, USA. Association for Computational Linguistics.
- Michael Schlichtkrull, Nedjma Ousidhoum, and Andreas Vlachos. 2023. [The intended uses of automated fact-checking artefacts: Why, how and who](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8618–8642, Singapore. Association for Computational Linguistics.
- Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. [Get your vitamin C! robust fact verification with contrastive evidence](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. [DeepSeekMath: Pushing the limits of mathematical reasoning in open language models](#). *arXiv preprint arXiv:2402.03300*.
- 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, New Orleans, Louisiana. Association for Computational Linguistics.
- James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. [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, Salar Mohtaj, Jing Yang, Veronika Solopova, and Vera Schmitt. 2025. [Comparing LLMs and BERT-based classifiers for resource-sensitive claim verification in social media](#). In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 281–287, Vienna, Austria. Association for Computational Linguistics.

- Juraj Vladika and Florian Matthes. 2023. [Scientific fact-checking: A survey of resources and approaches](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230, Toronto, Canada. Association for Computational Linguistics.
- David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. [Fact or fiction: Verifying scientific claims](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.
- David Wadden 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.
- David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. [SciFact-open: Towards open-domain scientific claim verification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Junjun Wang, Kunlong Chen, Zhaoqun Chen, Peng He, and Wenlu Zheng. 2025. [Winning Climate-Check: A multi-stage system with BM25, BGE-reranker ensembles, and LLM-based analysis for scientific abstract retrieval](#). In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 276–280, Vienna, Austria. Association for Computational Linguistics.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 2526–2547. Association for Computational Linguistics.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.
- 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, Chu-jie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.

A. Inter-Annotator Agreement And Annotation Process of Task 2

Since we use a multi-label annotation schema, we decompose the task into independent binary decisions per label. Table 5 reports agreement for all labels in the taxonomy.

Pairwise Cohen’s κ . Computed on flattened binary label vectors, κ ranges from 0.747 to 0.819 across all six annotator pairs (0.773–0.850 at the top-level), indicating substantial agreement.

Krippendorff’s α . At the top-level, α indicates substantial agreement on categories 0 and 1 ($\alpha \geq 0.67$), while showing moderate agreement on categories 2-5 ($\alpha = 0.54$ – 0.66). Among individual sub-narratives, the highest agreement was observed for 1_6 (“Sea level rise is exaggerated”; $\alpha = 0.795$), 3_2 (“Impacts are beneficial”; $\alpha = 0.688$), and 1_1 (“Ice isn’t melting”; $\alpha = 0.672$). Labels in the denial-of-cause family (2_1, 2_3, 2_5) and solutions-won’t-work family (4_2, 4_4) showed lower agreement ($\alpha = 0.26$ – 0.58), reflecting the greater interpretive difficulty of distinguishing closely related sub-narratives.

Positive Agreement. For rare labels, κ and α can be misleadingly low due to the prevalence paradox: agreement on label absence inflates expected agreement, deflating chance-corrected scores. We therefore also report Positive Agreement (PA), using the Dice coefficient over annotator pairs:

$$PA = \frac{2TP}{2TP + FP + FN}$$

For the most frequent disinformation labels (2_1, 5_1, 2_3), PA ranges from 0.57 to 0.59, confirming moderate positive-case agreement consistent with Krippendorff’s α .

Negative Agreement. Label 2_2 (“Human greenhouse gases are not causing climate change. It’s non-greenhouse gas human climate forcings (aerosols, land use)”) exhibited negative agreement ($\alpha = -0.003$), meaning annotators disagreed more than chance. It was applied 12 times across all annotators, never by more than one annotator on the same item. In each case, the claim discussed topics like deforestation or aerosols but stated factually accurate information; one annotator coded based on topic presence while the others correctly assigned 0_0. In 23 of the pairwise comparisons, the non-2_2 annotator assigned 0_0. The negative agreement thus reflects a systematic confusion between topic and narrative. After adjudication, 2_2 was retained in only one case.

Annotation Edge Cases. Three recurring annotation challenges during the annotation process: 1. Mention vs. endorsement: claims that reference a disinformation narrative but immediately question it (e.g., “Some argue that CO₂ isn’t the main driver of global warming. This is a controversial viewpoint, as the vast majority of scientists agree CO₂ plays a significant role in climate change.”) were labeled 0_0, as the text is *about* the narrative rather than an instance of it; 2. Disinformation intent in factually true statements: a claim such as “Jupiter’s weather is powered by a giant internal heat source. #Climate” is scientifically accurate, but the #Climate hashtag hints at an implicit natural-cycles framing (2_1); annotators disagreed on whether to code surface content or inferred intent, and such cases were resolved conservatively as 0_0; 3. Tone-dependent ambiguity: “Scientists are still working on the final numbers for the projected sea level rise” can be read as neutral reporting or as implying that climate science is unsettled (5_1); adjudicators independently converged on 5_1, interpreting “final” as doubting scientific consensus.

Summary. At the subnarrative level, agreement varies considerably. Labels with clear, observable indicators (e.g., 1_6: sea level claims; 1_1: ice/glacier claims) achieve substantial agreement, while labels requiring more interpretive judgment show moderate to fair agreement like within categories 3 (*impacts not bad*) and 4 (*solutions won’t work*). This pattern is consistent with the annotation evaluation of the original CARDS taxonomy, where top-level distinctions are more reliable than fine-grained ones Coan et al. (2021). The 16.6% of claims requiring adjudication were resolved by two authors who had not served as annotators, ensuring independence between the annotation and adjudication stages. Each adjudicator first reviewed the claim independently before a joint discussion, and all decisions were documented with reasoning.

B. Evaluation Metrics

Recall@K is defined as:

$$R@K = \frac{\# \text{gold evidentiary abstracts in top-K}}{\# \text{gold evidentiary abstracts for the claim}}$$

Because relevance judgments are incomplete, we additionally report Bpref, which does not assume exhaustive annotation (Buckley and Voorhees, 2004), and is defined as:

$$Bpref = \frac{1}{R} \sum_r \left(1 - \frac{|n \text{ ranked higher than } r|}{R} \right)$$

where R is the total number of judged relevant documents, r is a judged relevant document retrieved

Label	Narrative	N	Prev.	α	κ	PA
<i>0 — No disinformation narrative</i>						
0_0	No disinformation narrative detected	776	.715	.722	.722	.921
<i>1 — Global warming is not happening</i>						
1_0	Global warming is not happening (general)	12	.004	.175	.132	.133
1_1	Ice/permafrost/snow cover isn't melting	28	.017	.672	.676	.681
1_2	We're heading into an ice age/global cooling	23	.011	.556	.570	.575
1_3	Weather is cold/snowing	10	.004	.350	.313	.315
1_4	Climate hasn't warmed over the last decade(s)	36	.018	.544	.536	.543
1_5	Oceans are cooling/not warming	4	.002	.583	.610	.611
1_6	Sea level rise is exaggerated/not accelerating	22	.016	.795	.793	.797
1_7	Extreme weather isn't increasing/not linked to CC	22	.010	.427	.365	.370
1_8	Changed name from 'global warming' to 'climate change'	3	.002	.750	.738	.739
<i>2 — Human greenhouse gases are not causing climate change</i>						
2_0	Human GHGs are not causing CC (general)	14	.004	.080	.043	.044
2_1	It's natural cycles/variation	94	.051	.562	.561	.583
2_2	Non-GHG human forcings (aerosols, land use)	12	.003	-.003	-.003	.000
2_3	No evidence for GHG effect driving climate change	48	.028	.581	.578	.590
2_4	CO ₂ is not rising/ocean pH is not falling	3	.003	.800	.750	.750
2_5	Human CO ₂ emissions are miniscule	16	.006	.273	.277	.281
<i>3 — Climate impacts/global warming is beneficial/not bad</i>						
3_0	Climate impacts are beneficial/not bad (general)	14	.005	.230	.187	.189
3_1	Climate sensitivity is low/negative feedbacks	7	.003	.265	.281	.283
3_2	Species/ecosystems aren't impacted/are benefiting	22	.014	.688	.686	.690
3_3	CO ₂ is beneficial/plant food	30	.015	.568	.585	.591
3_4	It's only a few degrees (or less)	19	.006	.190	.175	.179
3_5	CC doesn't contribute to conflict/threaten security	8	.003	.131	.073	.074
3_6	CC doesn't negatively impact health	3	.002	.666	.694	.694
<i>4 — Climate solutions won't work</i>						
4_0	Climate solutions won't work (general)	2	.001	-.000	—	.000
4_1	Climate policies are harmful	13	.004	.122	.100	.103
4_2	Climate policies are ineffective/flawed	32	.013	.255	.254	.264
4_3	Too hard to solve (politically/economically/technically)	16	.006	.301	.297	.301
4_4	Clean energy/biofuels won't work	26	.010	.293	.260	.266
4_5	People need energy from fossil fuels/nuclear	9	.004	.398	.363	.365
<i>5 — Climate movement/science is unreliable</i>						
5_0	Climate movement/science is unreliable (general)	1	.000	.000	—	.000
5_1	Science is uncertain/unsound (data, methods, models)	70	.041	.551	.551	.570
5_2	Climate movement is alarmist/political/biased	26	.010	.238	.214	.221
5_3	Climate change is a conspiracy	6	.002	.249	.261	.262

Table 5: Per-label IAA sorted by top-level category. N = items where ≥ 1 annotator assigned the label; Prev. = prevalence; α = Krippendorff's alpha; κ = avg. pairwise Cohen's kappa; PA = positive agreement (Dice).

by the system, and n is a member of the first R judged non-relevant documents retrieved before r .

C. Ev²R Adaptation Details

C.1. Reference-based Component

To compute the reference-based component of Ev²R, we use the Gemini 2.5 Pro model (Comanici et al., 2025) to decompose gold and retrieved abstracts into atomic factual statements and evaluate their alignment. We adapt the prompt shared by Akhtar et al. (2024) to an instance from the Climat-

eCheck dataset, shown in Figure 6.¹⁶ For each retrieved abstract r and gold abstract g , the resulting F1 score is used as the reference-alignment score $F1(r, g)$. When multiple gold abstracts are available for a claim, we retain the abstract with the the maximum alignment score. If several gold abstracts achieve the same score, the first one is chosen. We cache the results across all submissions to prevent score variations for the same CAP.

¹⁶The full prompt is available at: https://github.com/ryabhmd/climatecheck/blob/master/automatic_eval/reference_based_prompt.txt

```

You will get as input a claim, a reference evidence and a predicted
evidence.
Please verify the correctness of the predicted evidence by comparing it
to the reference evidence, following these steps:
1. Break down the PREDICTED evidence in independent facts. Each fact
should be a separate sentence. Facts should not be specific to the
study, they must be generalized statements about climate science.
2. Evaluate each fact individually: is the fact supported by the
REFERENCE evidence? Do not use additional sources or background
knowledge.
3. Next, break down the REFERENCE evidence in independent facts. Each
fact should be a separate sentence. Facts should not be specific to
the study, they must be generalized statements about climate science.
4. Evaluate each fact individually: is the fact supported by the
PREDICTED evidence? Do not use additional sources or background
knowledge.
5. Finally summarise (1.) how many predicted facts are supported by
the reference evidence, (2.) how many reference facts are supported by
the predicted evidence.

Generate the output in form of a json as shown in the examples below.



---


Examples:
Claim: Antarctica is losing land ice at an accelerating rate,
according to satellite measurements.
Reference evidence: The Antarctic Ice Sheet is losing mass at an
accelerating rate, and playing a [...] causing a global sea-level
rise of about 3 m.
Predicted evidence: Greenland ice core records provide clear evidence
of rapid changes [...] its ice mass loss.

Output: {"facts in predicted evidence": "1. Greenland ice core
records provide clear evidence of rapid [...]", "facts count predicted
evidence": 5, "support predicted evidence": 2, "facts in reference
evidence": "1. Antarctic Ice Sheet is losing mass [...]", "facts
count reference evidence": 5, "support reference evidence": 1}



---


Input:
Claim: $claim
Reference evidence: $reference_evidence
Predicted evidence: $predicted_evidence

```

Figure 6: Prompt used for the Ev²R reference-based scorer adapted for the ClimateCheck shared task based on the prompt provided by Akhtar et al. (2024). Both the reference evidence and the retrieved evidence are decomposed into atomic facts representing generalised climate science statements before being assessed against each other.

C.2. Trained DeBERTa Classifier

To compute the proxy-reference component of Ev²R, as well as the automatic claim verification score for task 1.2, we train a DeBERTa-based classifier for 3-way NLI-style claim verification with the labels *Supports*, *Refutes*, and *NEI*.

The model is initialised from an existing classi-

fier,¹⁷ a DeBERTa-v3-Large checkpoint pre-trained on a diverse set of NLI datasets. We fine-tune it on a combined training set drawn from five datasets: FEVER (Thorne et al., 2018a), VitaminC (Schuster et al., 2021), HoVer (Jiang et al.,

¹⁷<https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli>

2020), AVeriTeC (Schlichtkrull et al., 2024), and our own ClimateCheck data. For ClimateCheck, we use a stratified 90/10 train/validation split, ignoring the official test split to prevent leakage. The combined training set comprises 488,018 examples and the validation set 77,843 examples.

Training is performed for 3 epochs on a single NVIDIA H100 GPU with a batch size of 32, a learning rate of 2×10^{-6} , linear learning rate scheduling with 200 warmup steps, and a maximum sequence length of 320 tokens. The best checkpoint is selected by evaluation accuracy on the combined validation set, evaluated every 2,000 steps. The selected checkpoint at step 26,000 (epoch 1.7) achieves a macro-F1 of 0.886 and an accuracy of 0.921 on the combined validation set, with per-class F1 scores of 0.955 (*supports*), 0.920 (*refutes*), and 0.782 (*NEI*). We make the model publicly available.¹⁸

D. Disinformation Narrative Classification Prompt

The prompt shown in Figure 7 was used for fine-tuning the baseline model for task 2.

You are an expert in detecting climate change related disinformation. You get a claim and your task is to classify the claim using the taxonomy codes provided.

Rules:

- If the claim aligns with narratives from the taxonomy, list applicable codes separated by a semicolon (;).
- If no disinformation is found, return '0_0'.

Taxonomy:
{taxonomy_str}

Claim: "{claim}"

Codes:

Figure 7: Prompt used for the baseline implementation of task 2.

E. Further Error Analyses

Figure 8 shows the confusion matrices of submitted systems from *berkbubus*, *gardlz*, and *ytoneva* on task 1.2. Table 6 displays the R@5 results for the two systems submitting for both tasks per narrative group based on the CARDS taxonomy. No notable differences in retrieval appear to exist, indicating that retrieval difficulty is narrative-agnostic.

¹⁸<https://huggingface.co/rausch/deberta-climatecheck-2463191-step26000>

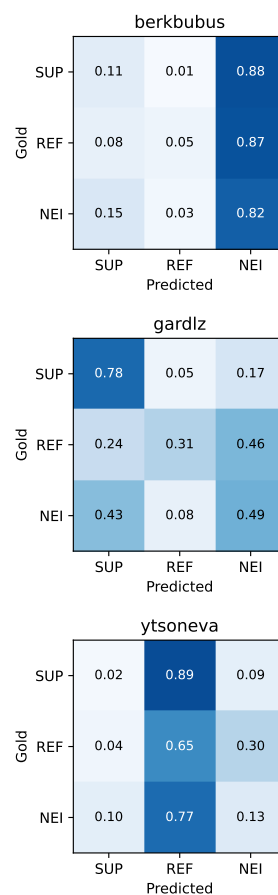


Figure 8: Confusion matrices for the *berkbubus*, *gardlz*, and *ytoneva* predictions on task 1.2, normalised by claim. SUP = Supports, REF = Refutes, NEI = Not Enough Information.

Narrative Group	R@5	N
<i>Baseline</i>		
0	0.992	124
1	0.933	15
2	0.938	16
3	1.000	6
4	1.000	6
5	1.000	11
<i>ClimateSense</i>		
0	0.992	124
1	1.000	15
2	1.000	16
3	1.000	6
4	1.000	6
5	0.909	11

Table 6: R@5 by narrative group for both systems. N indicates the number of claims per group. Results show minimal variance across narrative groups, indicating that retrieval difficulty is narrative-agnostic.