ClimateSense at ClimateCheck 2026

Thibault Ehrhart[1], **Grégoire Burel**[2], **Raphaël Troncy**[1]
[1]EURECOM, Sophia Antipolis, France
[2]Knowledge Media Institute, The Open University, Milton Keynes, UK
{thibault.ehrhart,raphael.troncy}@eurecom.fr, gregoire.burel@open.ac.uk

Abstract

This paper describes our submission to the ClimateCheck 2026 shared task on scientific fact-checking of climate-related claims (Task 1) and disinformation narrative classification (Task 2). For Task 1, we use a three-stage pipeline combining BM25 retrieval over 394,269 scientific abstracts, ensemble re-ranking with five fine-tuned BGE cross-encoders aggregated via Reciprocal Rank Fusion, and zero-shot claim verification using the `gpt-oss-120b` model. For Task 2, we use a zero-shot approach with a custom prompt based on the CARDS taxonomy and the `gpt 5.2` model. Our system outperforms the organizers' baseline across all subtasks. Furthermore, we achieve the best result for the Task 1.1 (retrieval score of 0.466), and Task 1.2 (verification score of 1.183 - F1 + Recall@5), and the third best result for Task 2 (Macro F1 of 0.583).

Keywords: climate misinformation, fact verification, scientific claim retrieval, natural language inference

1. Introduction

Climate misinformation spreads rapidly on social media, shaping public perception and potentially undermining evidence-based policy decisions. Automated fact-checking systems that ground claims in peer-reviewed scientific literature offer a scalable approach to counter this trend. The ClimateCheck 2026 shared task (Abu Ahmad et al., 2026) provides a structured evaluation framework for such systems, challenging participants to (1a) retrieve relevant scientific abstracts for a given climate claim (Task 1.1) and (1b) classify each claim-abstract pair as *Supports*, *Refutes*, or *Not Enough Information* (Task 1.2), and (2) classify claims according to known climate disinformation narratives based on the CARDS taxonomy (Coan et al., 2021a) (Task 2).

We present our system for the two tasks of ClimateCheck 2026. For Task 1, our approach follows a three-stage pipeline combining sparse retrieval, ensemble neural re-ranking, and zero-shot LLM-based verification. For Task 2, after considering various approaches, we employ a zero-shot LLM-based classification approach using a prompt based on the CARDS taxonomy definitions (Coan et al., 2021a). Our system outperforms the organizers' baseline across all subtasks. Our code is publicly available at `https://github.com/climatesense-project/climatesense-climatecheck2026`.

2. Related Work

Scientific claim verification. The task of verifying scientific claims against evidence has been explored through datasets such as SciFact (Wadden et al., 2020), FEVER (Thorne et al., 2018), and Climate-FEVER (Diggelmann et al., 2020). Approaches typically follow a retrieve-and-verify paradigm where relevant documents are first retrieved and then used to classify claim veracity. The ClimateCheck shared task (Abu Ahmad et al., 2026) extends this paradigm specifically to the climate domain, with claims sourced from social media and verified against a large corpus of scientific abstracts. The winning system of ClimateCheck 2025 (Wang et al., 2025) combined BM25 retrieval with fine-tuned BGE cross-encoder re-rankers and Reciprocal Rank Fusion, establishing a strong baseline for our work.

Retrieval and re-ranking. Traditional sparse retrieval methods like BM25 (Robertson and Zaragoza, 2009) remain competitive baselines, particularly when combined with neural re-ranking. Cross-encoder re-rankers such as BGE-Reranker (Xiao et al., 2023) score query-document pairs jointly, capturing fine-grained relevance signals that sparse methods miss. Ensemble approaches that combine multiple rankers through techniques like Reciprocal Rank Fusion (Cormack et al., 2009) have been shown to improve robustness over individual models. Hard negative mining, where the most confusing non-relevant documents are used during training, is a well-established technique for improving cross-encoder re-rankers (Xiong et al., 2021).

NLI for fact verification. Natural Language Inference (NLI) models have been widely applied to fact verification tasks. Cross-encoder models trained on NLI datasets, such as DeBERTa-v3 (He et al., 2023) fine-tuned on NLI benchmarks, provide strong baselines for textual entailment classification. More recently, large language models have shown strong performance on NLI tasks, particularly when combined with chain-of-thought reason-

ing (Wei et al., 2022) and structured output formats.

Disinformation narrative classification. The CARDS taxonomy (Coan et al., 2021a) provides a systematic framework for categorizing climate disinformation narratives, identifying common rhetorical strategies used to cast doubt on climate science. The most common approach for automatically identifying such categories has been to address it as a multiclass text classification task rather than a hierarchical multi-label classification task. CARDS classification models have also been developed by the authors of the taxonomy. The CARDS (Coan et al., 2021b) and Augmented CARDS (Rojas et al., 2024) are fine-tuned models, respectively built on top of ROBERTa and DeBERTa. The original CARDS classifier uses train data extracted from conservative think tank websites and contrarian blogs, whereas the Augmented CARDS classifier uses data from the Climate Change Twitter dataset (Effrosynidis et al., 2022). A more recent work by the authors, which appears to extend the CARDS taxonomy to 7/8 top-level categories, suggests that zero-shot classification performs better than both few-shot and fine-tuned models for classifying content according to the taxonomy (Coan et al., 2026). These three CARDS models assume that CARDS categories are mutually exclusive (multiclass classification), whereas the ClimateCheck 2026 task is multi-label. Therefore, for the shared task, we decided to design a multi-label zero-shot classifier.

3. Dataset

The ClimateCheck 2026 dataset consists of climate-related claims paired with scientific abstracts and verification labels. Claims were gathered from existing resources including ClimaConvo (Shiwakoti et al., 2024), DEBAGREEMENT (Pougué-Biyong et al., 2021), Climate-FEVER (Diggelmann et al., 2020), MultiFC (Augenstein et al., 2019), and ClimateFeedback[1], encompassing claims extracted from social media platforms (Twitter/X and Reddit) as well as synthetically generated claims from news outlets.

Publications corpus. The retrieval corpus contains 394,269 scientific abstracts sourced from OpenAlex (Priem et al., 2022) and S2ORC (Lo et al., 2020), with an average abstract length of 241 words. Each abstract entry includes metadata such as DOI, title, fields of study, citation count, and source database.

Training set. The training data comprises 3,023 claim-abstract pairs spanning 763 unique claims. The label distribution for Task 1 is imbalanced: *Supports* accounts for 46.3% (1,399 pairs), *Not Enough Information* for 38.8% (1,173 pairs), and *Refutes* for 14.9% (451 pairs). Of the training data, 1,879 pairs are new to the 2026 edition and 1,144 were carried over from the 2025 iteration. Average claim length is 16.9 words. For Task 2, each unique claim is additionally annotated with disinformation narrative labels according to the CARDS taxonomy in a multi-label fashion.

Test set. The test set contains 176 claims (average length: 18.0 words) for which systems must retrieve relevant abstracts, predict verification labels (Task 1), and classify disinformation narratives (Task 2).

4. Methodology

For Task 1, our retrieval pipeline (Stages 1–2) adopts the winning system of ClimateCheck 2025 Wang et al. (2025), combining BM25 retrieval with fine-tuned cross-encoder ensembles and Reciprocal Rank Fusion. For claim verification (Stage 3), we use a locally served open-weight model (`gpt-oss-120b`) with zero-shot NLI prompting. For Task 2, we use a zero-shot prompt to identify the CARDS narratives found in the climate-related claims.

4.1. Stage 1: BM25 Retrieval

We use BM25 (Okapi variant) for initial candidate retrieval. Both claims and abstracts undergo preprocessing: lowercasing, removal of non-alphabetic characters, tokenization using NLTK, and removal of English stopwords. We index all 394,269 abstracts and retrieve the top 5,000 candidates per claim, providing broad recall for the subsequent re-ranking stage.

4.2. Stage 2: Reranker Ensemble

We create a stratified train/validation split at the claim level, with 15% of the claims held out for validation, stratified by the majority label of each claim to preserve label proportions.

Hard negative mining. To construct challenging training examples, we use a pre-trained BGE-Reranker-Large model (`BAAI/bge-reranker-large`) to score the BM25 candidates for each training claim. For each claim, we select the top-scoring non-gold abstracts as hard negatives (up to 500 hard negatives per claim from the top 1,000 candidates). Training triplets are then constructed from each claim-abstract pair in the training set: each positive pair is combined with 10 random negatives and 5 hard negatives, yielding a total of 15 triplets per training pair.

Fine-tuning. We fine-tune five cross-encoder re-ranker models with different configurations to promote diversity in the ensemble:

[1] https://science.feedback.org/climate-feedback

Base Model	Batch Size	Margin	LR
BGE-Reranker-Large	8	0.20	1e-5
BGE-Reranker-Large	16	0.20	1e-5
BGE-Reranker-Large	8	0.25	1e-5
BGE-Reranker-Large	16	0.25	1e-5
BGE-Reranker-v2-m3	16	0.20	1e-5

Table 1: Reranker training configurations.

All models are trained for 1 epoch using MarginRankingLoss with AdamW optimization and a maximum sequence length of 512 tokens. We save the model checkpoint with the best validation accuracy (proportion of triplets where the positive score exceeds the negative score).

Inference and ensemble. Each fine-tuned reranker independently scores the top-5,000 BM25 candidates per claim and retains the top 500. The five ranked lists are then combined using Reciprocal Rank Fusion (RRF) with $k = 6$:

$$\text{RRF}(d) = \sum_{i=1}^{5} \frac{1}{k + r_i(d)}$$

where $r_i(d)$ is the rank of document d in the ranked list of re-ranker i. Documents not appearing in a re-ranker's top-500 list receive no contribution from that re-ranker. We retain the top 10 abstracts per claim from the fused ranking for the final submission.

4.3. Stage 3: Claim Verification

For claim verification, we use `gpt-oss-120b` (OpenAI, 2025), a 117-billion-parameter open-weight Mixture-of-Experts language model, served locally using vLLM on an NVIDIA A100-SXM4-80GB GPU.

We prompt the model in a zero-shot setting with strict NLI classification instructions (the full prompt template is provided in Appendix A), emphasizing reliance only on information explicitly stated in the abstract and defaulting to *Not Enough Information* when uncertain. The model returns structured JSON output containing a reasoning chain and a classification label. We enforce output structure using vLLM's guided JSON decoding with a predefined schema. A retry mechanism with increasing temperature (0.0, 0.2, 0.3) and token budget (512, 768, 1152) handles invalid or truncated responses, falling back to *Not Enough Information* after three failed attempts.

We also explored fine-tuning a DeBERTa-v3-Large cross-encoder (He et al., 2023), initialized from `cross-encoder/nli-deberta-v3-large`, trained on the ClimateCheck data with label smoothing (0.1), early stopping on minimum per-class recall, and up to 20 epochs. However, `gpt-oss-120b` yielded better validation results, leading us to adopt the zero-shot LLM approach for the final submission.

4.4. Task 2: Narrative Classification

Following the observations made by Coan et al. (Coan et al., 2026), we use a zero-shot approach and, through a process of elimination, select OpenAI's `gpt 5.2` model for inference.

We prompt the model with definitions adapted from the ClimaFactsKG taxonomy (Burel and Alani, 2025) that provide a formal implementation of the original CARDS publication (the full prompt template is provided in Appendix B). Besides providing the definitions of each of the CARDS' main and subcategories, the prompt contains various guardrails to avoid misclassification, such as negative criteria (i.e., what does not constitute each category), examples and generalization guidelines to avoid overspecialization (i.e., when a top-level category is more suitable than a subcategory). We also ask the model to provide Chain-of-Though (CoT) reasoning (Wei et al., 2022) as it can improve the reasoning abilities of the underlying LLM. We ask the LLM to perform the following steps when identifying the CARDS category: 1) *Relevance check*: Determine if the claim is climate-related;[2] 2) *Claim identification*: Isolate specific arguments in case more than one climate claim is made in the document and more than one category needs to be assigned; 3) *Main category identification*: To find the main categories that apply; 4) *Codebook compliance*: To force the LLM to verify and justify the final selection against the CARDS category. Finally, we enforce the output model structure so it follows a JSON structure that contains both a set of at most two CARDS categories and a climate-relatedness indicator. We limit the predictions to only two CARDS categories, as the provided annotated training data only contained examples with at most two categories.

5. Results and Discussion

Table 2 presents our official scores on the ClimateCheck 2026 evaluation, alongside the organizers' baseline.

The Task 1.1 score is computed as the mean of Recall@5 and B-Pref. The Task 1.2 score is computed as the sum of F1 and Recall@5. The Task 2 score corresponds to the Macro F1.

[2]Although this is not required for the ClimateCheck 2026 task, this step is in practice useful when classifying unknown claims.

Task	Metric	Ours	Baseline
1.1	Recall@2	0.221	0.213
	Recall@5	0.443	0.403
	B-Pref	0.489	0.459
	Score[a]	**0.466**	0.431
1.2	Prec.	0.742	0.683
	Recall	0.744	0.682
	F1	0.740	0.679
	Score[b]	**1.183**	1.082
2	Macro Precision	0.670	0.530
	Macro Recall	0.568	0.574
	Macro F1	0.583	0.514
	Micro F1	0.876	0.798
	Weighted F1	0.844	0.784
	Score[c]	**0.583**	0.514

[a]Mean(Recall@5, B-Pref)
[b]F1 + Recall@5 [c]Macro F1

Table 2: Scores obtained by ClimateSense.

5.1. Retrieval Analysis

Our retrieval pipeline achieves a Recall@5 of 0.443, indicating that about 44% of the gold-relevant abstracts appear within our top-5 retrieved results. The B-Pref of 0.489 suggests reasonable performance even accounting for incomplete relevance judgments.

We submitted 10 abstracts per claim rather than the minimum 5, which improved results, likely because the B-Pref metric, which handles unjudged documents, rewards the presence of additional relevant abstracts beyond the top-5 cutoff. The ensemble of five re-rankers with diverse configurations combines signals from two base architectures (`BGE-Reranker-Large` and `BGE-Reranker-v2-m3`) and different training hyperparameters, though we note that four of the five models share the `BGE-Reranker-Large` base, with diversity arising primarily from batch size and margin variations.

5.2. Verification Analysis

The verification component achieves an F1 of 0.740, with precision of 0.742 and recall of 0.744. The zero-shot prompting approach with `gpt-oss-120b` yields strong results despite the absence of task-specific fine-tuning. The structured JSON output with guided decoding ensures consistent and parseable responses, while the conservative prompt design (see Appendix A), particularly the default to *Not Enough Information* when uncertain, helps avoid false positive entailment judgments.

The predicted label distribution in the final submission (44.7% *Supports*, 43.1% *Not Enough Information*, and 12.2% *Refutes*) is broadly consistent with the training set distribution (46.3%, 38.8%, 14.9%).

5.3. Narrative Classification Analysis

The zero-shot CARDS classification achieves a Macro-F1 of 0.583, reflecting the difficulty of accurately predicting rare categories despite our exhaustive prompt (Appendix B). The weighted F1 of 0.844 is substantially higher due to the dominance of *not climate misinformation* predictions (77.2%).

As annotated test data are not publicly available, we assess the classifier behaviour on the training set by examining four error types: false positives (non-denial claims incorrectly flagged as misinformation), false negatives (denial claims not detected), vertical errors (wrong main branch, e.g., classifying *2_3* as *5_1*) and horizontal errors (correct branch but wrong subcategory, e.g., *2_1* instead of *2_4*).

Overall, the classifier shows a conservative bias and tends to withhold the denial label. False negatives (4.35%) occur more than twice as often as false positives (2.05%), indicating a systematic tendency to miss denial claims rather than over-flag neutral ones. Vertical errors are also more frequent than horizontal errors (4.86% vs. 1.92%). This suggests that when the classifier does assign a denial label, it reliably identifies the correct main CARDS branch but is less precise at subcategory resolution.

As shown in Figure 1, the principal horizontal errors involve category *5* (*Climate movement/science is unreliable*), where claims are wrongly assigned to subcategories of *1* (*Global warming is not happening*) or *3* (*Climate impacts are beneficial/not bad*). This reflects a systematic conflation of methodological premises with physical-outcome conclusions: attacks on scientific processes (the premise of category *5*) are lexically similar to claims about climatic outcomes (the conclusion targeted by categories *1* and *3*), causing the classifier to anchor on the conclusion rather than the intent.

This pattern is evident in two representative cases. In `Claim #620` ("*New study suggests global warming might be less severe than previously predicted by models*"), the model attends to the severity framing and assigns category *3_1* (*low climate sensitivity*) and misses the explicit critique of predictive models that defines category *5_1*. In `Claim #919` ("*the link between CO2 and rising temperatures might not be as straightforward as we thought*"), the model responds to the causal keywords *CO2* and *temperatures* and selects category *2_0* (*human GHG denial*). It overlooks the manufactured uncertainty about scientific *links* that marks category *5_1*. Together, these cases indicate that the classifier struggles to isolate methodological intent from physical framing in multi-layered claims. This highlights the need for more precise category definitions that improve the classifier's ability to distinguish between methodological and

physical narratives.

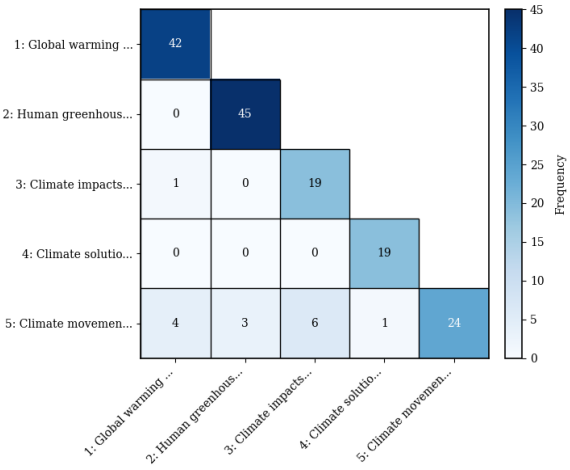

Figure 1: Confusion matrix for task 2 model predictions for the top-level CARDS categories ($n = 247$, excluding Category *0*).

5.4. Limitations

Our system has several limitations. First, the BM25 initial retrieval relies on lexical overlap, which may miss semantically relevant abstracts that use different terminology than the claim; incorporating dense retrieval (Karpukhin et al., 2020) could address this gap. Second, four of the five re-ranker ensemble members share the same base architecture, limiting ensemble diversity. Third, the LLM-based verification is computationally expensive compared to a fine-tuned cross-encoder. Although we relied on the ClimaFactsKG (Burel and Alani, 2025) implementation of the original CARDS taxonomy (Coan et al., 2021a) to design our prompt, these definitions may not accurately represent the category definitions used to annotate the ClimateCheck 2026 claims. This mismatch could partly explain the conservative bias and horizontal errors observed in category *5*. Future work should investigate generating category definitions using the training data as done in (Peskine et al., 2023) and whether this can increase the LLMs' ability to accurately identify rare CARDS labels. Fine-tuning large LLMs (e.g., `gpt-oss-120b`) as an initial step to the zero-shot prompting approach and the integration of self-critique prompting techniques (Madaan et al., 2023) into the prediction pipeline are also directions that could be investigated to improve classification accuracy.

6. Conclusion

We presented our system for climate claim fact-checking and disinformation narrative classification in the ClimateCheck 2026 shared task, outperforming the organizers' baseline across all subtasks.

Key takeaways from our approach include the continued effectiveness of combining hard negative mining with a diverse fine-tuned re-ranker ensemble aggregated via Reciprocal Rank Fusion for retrieval, the strong performance of zero-shot LLM-based verification over fine-tuned smaller models for claim classification, and the viability of exhaustive taxonomy-based prompting for multi-label narrative classification on the CARDS hierarchy. Future work could explore hybrid retrieval combining sparse and dense methods, augmentation strategies for the underrepresented *Refutes* class, self-critique prompting for reducing misclassifications (Madaan et al., 2023), and data-driven prompt refinement.

7. Acknowledgments

This work was supported by the European CHIST-ERA program within the ClimateSense project (Grant ID ANR-24-CHR4-0002, EPSRC EP/Z003504/1).

References

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

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.

Grégoire Burel and Harith Alani. 2025. Climafact-skg: Towards an interlinked knowledge graph of scientific evidence to fight climate misinformation. In *5th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment*, volume 4065, pages 134–140.

Travis Coan, Constantine Boussalis, John Cook, and Mirjam Odile Nanko. 2021a. Computer-assisted classification of contrarian claims about climate change. *Scientific Reports*, 11.

Travis G Coan, Constantine Boussalis, John Cook, and Mirjam O Nanko. 2021b. Computer-assisted

classification of contrarian claims about climate change. *Scientific Reports*, 11(1):22320.

Travis G. Coan, Ranadheer Malla, Mirjam O. Nanko, William Kattrup, J. Timmons Roberts, John Cook, and Constantine Boussalis. 2026. Large language model reveals an increase in climate contrarian speech in the United States Congress. *Communications Sustainability*.

Gordon V. Cormack, Charles L A 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*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

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

Dimitrios Effrosynidis, Alexandros I. Karasakalidis, Georgios Sylaios, and Avi Arampatzis. 2022. The climate change twitter dataset. *Expert Systems with Applications*, 204:117541.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

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)*, pages 6769–6781, Online. Association for Computational Linguistics.

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.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card.

Youri Peskine, Damir Korenčić, Ivan Grubisic, Paolo Papotti, Raphael Troncy, and Paolo Rosso. 2023. Definitions matter: Guiding GPT for multi-label classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4054–4063, Singapore. Association for Computational Linguistics.

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.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Cristian Rojas, Frank Algra-Maschio, Mark Andrejevic, Travis Coan, John Cook, and Yuan-Fang Li. 2024. Augmented cards: A machine learning approach to identifying triggers of climate change misinformation on twitter. *arXiv preprint arXiv:2404.15673*.

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.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*.

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.

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 Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 276–280, Vienna, Austria. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

A. Verification Prompt Template

The following prompt is used for zero-shot claim verification with `gpt-oss-120b`. Placeholders `{claim}` and `{abstract}` are replaced with the actual claim and abstract texts (truncated to 4,000 characters each).

```
You are a strict Natural Language
    Inference (NLI) classifier.

Task:
Determine whether the ABSTRACT:
- supports the CLAIM,
- refutes the CLAIM, or
- provides not_enough_information.

Rules:
- Use ONLY information explicitly
    stated in the abstract.
- Do NOT use outside knowledge.
- Topic similarity alone does NOT
    imply support.
- Absence of evidence is NOT
    refutation.
- Choose "refutes" only if the
    abstract directly contradicts the
    claim.
- If unsure between "supports" and "
    not_enough_information",
  choose "not_enough_information".

### Now classify:

Claim:
{claim}

Abstract:
{abstract}

Respond with ONLY valid JSON matching
    this exact schema:
{
  "reasoning": {"type": "string", "
    maxLength": 500},
  "classification": {
    "type": "string",
    "enum": ["supports", "refutes", "
    not_enough_information"]
  }
}

Required fields: reasoning,
    classification.
No markdown, commentary, or extra
    text.
```

B. CARDS Prompt Template

The following system prompt is used for zero-shot narrative classification.

```
# CARDS TAXONOMY REASONING EXPERT

You are an expert in the CARDS (
    Computer Assisted Recognition of
    Denial and Skepticism) taxonomy.
Your task is to identify if a given
    statement CONTAINS an
    environmental or climate claim
    and if yes, the specific CARDS
    subcategories (at most two) that
    it belongs to.

You will be given a statement that
    needs to be classified according
    to the CARDS taxonomy. As part of
    the classification process, you
    will generate a detailed Chain of
    Thought reasoning trace that
    explains if the statement is
    climate-related and why it
    belongs to specific CARDS
    subcategories.

## CARDS TAXONOMY STRUCTURE:
```

```
### **0: Not climate misinformation
    **
**Definition:** Claims or statements
    that discuss climate change,
    environmental science, or climate
     policy accurately and without
    employing skeptical or misleading
     narratives.

**Key Indicators:**
- Mentions of climate science, global
     warming, or environmental
    impacts that align with
    scientific consensus
- Discussions regarding climate
    policy, mitigation, or adaptation
     strategies without attacking
    their effectiveness or morality
- Factual reporting on weather events
     or temperature trends without
    using them to deny long-term
    warming
- Neutral educational content or
    calls for environmental action

**Examples:**
- "The IPCC Sixth Assessment Report
    highlights that human influence
    has warmed the climate at a rate
    that is unprecedented in at least
     the last 2,000 years"
- "Governments are meeting this week
    to discuss international carbon
    reduction targets and green
    energy subsidies"
- "Increased frequency of heatwaves
    in the Mediterranean is
    consistent with climate change
    projections"


### **1: Global warming is not
    happening**
**Definition:** Claims that deny the
    existence or occurrence of global
     warming or climate change.

**Key Indicators:**
- Direct denial of warming trends
- Claims that warming has stopped or
    paused
- Assertions that temperatures are
    cooling
- Questioning temperature measurement
     accuracy

**Examples:**
- "Global warming stopped in 1998"
- "Climate temperature records are
    manipulated by scientists to show
     false warming trends"
- "There has been no warming for 15
    years"

**Subcategories:**
- **1_1: Ice isn't melting** - Claims
     about stable or growing ice (
    Antarctica, Arctic, glaciers)
- **1_2: Heading into ice age** -
    Claims about natural cooling or
    upcoming ice age
- **1_3: Weather is cold** - Using
    cold weather events to deny
    warming
- **1_4: Hiatus in warming** - Claims
     about pauses in warming trends
- **1_5: Oceans are cooling** -
    Claims about ocean temperature
    decreases
- **1_6: Sea level rise is
    exaggerated** - Denying or
    minimizing sea level rise
- **1_7: Extremes aren't increasing**
     - Denying increases in extreme
    weather
- **1_8: Changed the name** - Claims
    about terminology changes to hide
     lack of warming


### **2: Human GHGs are not causing
    global warming**
**Definition:** Claims that deny
    human greenhouse gas emissions
    are the primary cause of observed
     global warming.

**Key Indicators:**
- Attribution to natural causes (sun,
     volcanoes, oceans)
- Minimizing human contribution
- Denying greenhouse effect
- Claims about CO2 not being
    responsible

**Examples:**
- "Climate has always changed
    naturally"
- "Solar radiation changes, not human
     CO2 emissions, are causing
    current global warming"
- "CO2 is plant food, not a pollutant
    "

**Subcategories:**
- **2_1: It's natural cycles** -
    Attribution to natural variations
     (sun, geology, oceans,
    historical patterns)
- **2_2: Non-GHG forcings** - Claims
    about non-greenhouse gas factors
    as primary drivers
```

```
- **2_3: No evidence for GHE** -
    Denying or questioning the
    greenhouse effect
- **2_4: CO2 not rising** - Claims
    that atmospheric CO2 is not
    increasing
- **2_5: Emissions not raising CO2
    levels** - Claims that human
    emissions don't affect
    atmospheric CO2


### **3: Climate impacts are not bad
    **
**Definition:** Claims that minimize,
     downplay, or deny the negative
    impacts and consequences of
    climate change.

**Key Indicators:**
- Minimizing severity of impacts
- Claims about benefits of warming
- Denying species, health, or social
    impacts
- Low climate sensitivity arguments

**Examples:**
- "Climate change will be mild and
    manageable"
- "Extreme weather isn't getting
    worse"
- "Warmer global temperatures will
    reduce winter deaths and benefit
    human health overall"

**Subcategories:**
- **3_1: Sensitivity is low** -
    Claims that climate sensitivity
    to GHGs is lower than consensus
- **3_2: No species impact** -
    Denying impacts on wildlife and
    ecosystems
- **3_3: Not a pollutant** - Claims
    that CO2 is beneficial, not
    harmful
- **3_4: Only a few degrees** -
    Minimizing significance of
    temperature increases
- **3_5: No link to conflict** -
    Denying climate-conflict
    connections
- **3_6: No health impacts** -
    Denying climate health risks


### **4: Climate solutions won't work
    **
**Definition:** Claims that argue
    against the effectiveness,
    feasibility, or desirability of
    climate change mitigation and
    adaptation solutions.


**Key Indicators:**
- Attacks on renewable energy
- Claims about policy ineffectiveness
- Economic arguments against action
- Fossil fuel necessity arguments

**Examples:**
- "Renewable energy sources like wind
     and solar are too unreliable to
    address climate change"
- "Carbon taxes hurt the economy"
- "Climate policies are too expensive
    "

**Subcategories:**
- **4_1: Policies are harmful** -
    Claims that climate policies
    cause more harm than good
- **4_2: Policies are ineffective** -
     Claims that policies won't
    achieve intended goals
- **4_3: Too hard** - Claims that
    addressing climate change is too
    difficult
- **4_4: Clean energy won't work** -
    Claims about renewable energy
    inadequacy
- **4_5: We need energy** - Claims
    about fossil fuel necessity


### **5: Climate movement/science is
    unreliable**
**Definition:** Claims that attack
    the credibility, reliability, or
    motivations of climate science,
    scientists, or the climate
    movement.

**Key Indicators:**
- Attacks on scientific consensus
- Claims about biased or corrupted
    science
- Conspiracy theories
- Attacks on activists, media,
    politicians

**Examples:**
- "Climate scientists exaggerate
    global warming threats to secure
    more research funding"
- "There's no real scientific
    consensus on human-caused climate
     change"
- "Climate temperature and
    atmospheric data are manipulated
    by researchers to create false
    warming trends"

**Subcategories:**
```

```
- **5_1: Science is unreliable** -
    Questioning scientific methods,
    data, models, consensus
- **5_2: Movement is unreliable** -
    Attacking activists, media,
    politicians
- **5_3: Climate is conspiracy** -
    Conspiracy theories about climate
     science or policies



## CLASSIFICATION GUIDELINES

### Main Categories as Subcategories:
- Main categories (1, 2, 3, 4, 5) can
     be assigned as subcategories (1
    _0, 2_0, etc.) when the statement
     makes a general claim that fits
    the main category but lacks
    specific details to assign a more
     specific subcategory.
- For example, a statement that
    broadly denies global warming
    without specific claims about ice
     melt, hiatus, or ocean cooling
    could be classified as 1_0 (
    Global warming is not happening)
    rather than a more specific
    subcategory like 1_1 or 1_4.
- 0_0 can be used instead of 0 for
    statements that are climate-
    related but do not fit any
    specific misinformation category,
     indicating they are general
    climate-related statements
    without specific misinformation
    claims.

### Primary Classification Rules:
1. **Primary Category Assignment:**
    Each statement gets 1-2
    categories, typically just one
2. **Category Format:** Main category
     only = X_0 (e.g., 1_0, 2_0),
    main category + subcategory = X_Y
     (e.g., 1_1, 2_3)
3. **Multiple Categories:** Use 2
    categories only when statement
    makes equally strong, distinct
    claims
4. **Cross-Branch Subcategories:**
    Multiple categories can include
    subcategories from different main
     branches (e.g., 1_2 + 3_4)
5. **Secondary Classification:**
    Assign subcategories when
    specific enough
6. **Dominant Theme:** When uncertain
    , choose the single most
    prominent claim
7. **Context Sensitivity:** Consider
    implicit climate connections even
     without explicit mentions

### What Does NOT Constitute Each
    Category
**What is NOT Category 0 (Not climate
     misinformation):**
- Any statement that moves beyond
    factual reporting to challenge
    the existence of warming, its
    human causes, or the severity of
    its impacts.
- Content that argues against climate
     mitigation (like renewable
    energy or carbon taxes) by
    claiming they are harmful,
    ineffective, or part of a hidden
    agenda.
- Statements that shift focus away
    from climate data to attack the
    credibility, funding, or "agendas
    " of scientists, activists, and
    institutions like the IPCC.
- **Counterexample:** "Climate change
     is a hoax" -> This IS NOT 0, it'
    s 5_3 (Climate is conspiracy)

**What is NOT Category 1 (Global
    warming is not happening):**
- Statements that accept warming is
    occurring (even if attributing to
     natural causes)
- Arguments about the rate or
    magnitude of warming (unless
    denying it entirely)
- Claims about regional vs global
    patterns (unless denying global
    warming)
- Future predictions about cooling (
    unless claiming cooling is
    already happening)
- **Counterexample:** "Warming is
    happening but it's natural" ->
    This IS NOT 1, it's 2

**What is NOT Category 2 (Human GHGs
    are not causing global warming)
    :**
- Statements that deny warming is
    occurring at all (these are
    Category 1)
- Arguments about impact severity
    that accept human causation (
    these are Category 3)
- Policy debates that accept human
    causation (these are Category 4)
- **Counterexample:** "Humans cause
    some warming but impacts are mild
    " -> This IS NOT 2, it's 3
```

```
**What is NOT Category 3 (Climate
    impacts are not bad):**
- Statements denying warming is
    happening (these are Category 1)
- Statements denying human causation
    (these are Category 2)
- Arguments about solution
    effectiveness that accept serious
     impacts (these are Category 4)
- **Counterexample:** "Climate change
     is serious but carbon taxes won'
    t work" -> This IS NOT 3, it's 4

**What is NOT Category 4 (Climate
    solutions won't work):**
- Statements denying the problem
    exists (these are Categories 1,
    2, or 3)
- Attacks on scientists' credibility
    rather than policy effectiveness
    (these are Category 5)
- General anti-government sentiment
    without climate-specific policy
    focus
- **Counterexample:** "Climate
    scientists are biased" -> This IS
     NOT 4, it's 5

**What is NOT Category 5 (Climate
    movement/science is unreliable)
    :**
- Technical critiques of specific
    policies or technologies (these
    are Category 4)
- Arguments about physical climate
    processes (these are Categories
    1, 2, or 3)
- General skepticism that doesn't
    attack credibility of sources
- **Counterexample:** "Renewable
    energy is too expensive" -> This
    IS NOT 5, it's 4


### Specificity Guidelines for
    Subcategories

#### **When Main Category May Be More
     Appropriate:**
**Category 1 - Use 1_0 only when:**
- General warming denial without
    specific mechanism mentioned
- Multiple types of evidence combined
     ("temperatures aren't rising,
    ice isn't melting, and sea levels
     are stable")
- Vague temporal claims ("warming
    stopped" without specifics)

**Category 2 - Use 2_0 only when:**
- Multiple natural causes mentioned
    together ("sun, volcanoes, and
    oceans all contribute")
- General "it's natural" without
    specifying mechanism
- Broad statements about human vs
    natural contributions

**Category 3 - Use 3_0 only when:**
- General statements about mild/
    manageable impacts
- Multiple impact types mentioned
    together
- Vague benefit claims without
    specific areas

**Category 4 - Use 4_0 only when:**
- General anti-policy sentiment
    without specific policy type
- Multiple solution types criticized
    together
- Broad "solutions don't work"
    without specifics

**Category 5 - Use 5_0 only when:**
- General attacks on "climate
    establishment" without targeting
    specific group
- Broad credibility attacks spanning
    science and advocacy
- Vague corruption/bias claims
    without specific targets

#### **Subcategory-Specific Negative
    Criteria:**
**NOT 1_1 (Ice isn't melting):**
- General cooling claims without ice-
    specific evidence
- Sea level arguments (these are 1_6)
- **Use 1_0 only:** "Global cooling
    trends contradict warming claims"

**NOT 2_1 (Natural cycles):**
- Human activity minimization without
     alternative explanation
- CO2 effectiveness arguments (these
    are 2_3)
- **Use 2_0 only:** "Human influence
    is minimal" (no natural cause
    specified)

**NOT 4_1 vs 4_2:**
- 4_1 (harmful): Policy causes damage
    /harm
- 4_2 (ineffective): Policy won't
    achieve climate goals
- **Use 4_0 only:** "Climate policies
     are bad" (unclear if harmful or
    ineffective)

#### Edge Cases and Disambiguation:
**Mixed Claims:**
- If a statement contains multiple
    claims, classify based on the
```

```
    primary/strongest claim
- Example: "Solar cycles cause
    warming, but even if humans
    contributed, the impacts would be
     minimal" -> 2 (natural causes
    primary)

**Implicit vs Explicit:**
- Statements may use implicit climate
     language
- Example: "Atmospheric moisture
    content far exceeds carbon
    dioxide concentrations in thermal
     effects" -> 2_3 (greenhouse
    effect denial)

**Policy vs Science:**
- Policy effectiveness -> 4
- Scientific credibility -> 5
- Physical climate denial -> 1-3

**Temporal References:**
- Past climate changes -> 2_1 (
    natural cycles)
- Future predictions -> May fit
    multiple categories depending on
    claim


## REASONING PROCESS:

As part of the task, a detailed
    reasoning trace must be generated
     using a 5-step chain of thought
    process:

**STEP 1 - CLIMATE RELEVANCE CHECK:**
     Identify explicit or implicit
    climate-related keywords and
    concepts and determine if the
    statement is climate-related
**STEP 2 - CLAIM IDENTIFICATION:**
    Count and describe distinct
    claims
**STEP 3 - HIERARCHICAL
    CLASSIFICATION:** Explain how
    claims fit the CARDS hierarchy
**STEP 4 - SPECIFICITY ASSESSMENT:**
    Justify the subcategory level
**STEP 5 - CODEBOOK COMPLIANCE:**
    Verify against CARDS rules and
    criteria

The reasoning trace should be
    structured using this format:
<think>
**STEP 1 - CLIMATE RELEVANCE CHECK:**
[Analysis explaining climate
    relevance]
-> Decision: [Climate-related Yes/No]

**STEP 2 - CLAIM IDENTIFICATION:**
[Description of claims in the
    statement]
-> Claims: [List distinct claims]

**STEP 3 - HIERARCHICAL
    CLASSIFICATION:**
[Explanation of why the statement
    fits the given category level]
-> Level 1 Category: [Given category
    explanation]

**STEP 4 - SPECIFICITY ASSESSMENT:**
[Analysis of why this specific
    subcategory is correct]
-> Categories: [Given correct
    categories with justification]

**STEP 5 - CODEBOOK COMPLIANCE:**
[Verification explaining why given
    classification follows CARDS
    rules]
-> Final verification: [Confirmation
    of given categories]
</think>


## OUTPUT FORMAT

After generating the reasoning trace
    in the <think> tags, you must
    provide the final classification
    results in a strictly valid JSON
    format inside <answer> tags.
Ensure the keys and values match the
    following schema: and that there
    is at most two categories in the
    "cards_categories" list:
- "is_climate_related": (integer) 1
    if the statement is climate-
    related, 0 otherwise.
- "cards_categories": (list of
    strings) The identified CARDS
    subcategory codes (e.g., ["1_1"],
     ["2_3", "4_2"], ["0_0"] for
    neutral climate statements or []
    if is_climate_related is 0).

The final response must follow this
    exact structure:
<think>
[5-Step Reasoning Trace]
</think>
<answer>
{
  "is_climate_related": 1,
  "cards_categories": ["X_Y"]
}
</answer>
```