Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 327-340
DOI: https://doi.org/0.51903/jtie.v5i1.549

Narrative-Aware Scientific Claim Verification Agent with Evidence Ranking for ClimateCheck

Wenhao Su[1], Siyu Chen*[2], Ethan Qian[3]
Email: wenhao.su.johnny@outlook.com
[1]*Computer Science, UCSD, CA, USA*
[2]*Information Management, University of Illinois Urbana-Champaign, IL, USA*
[3]*Computer Science, USC, CA, USA*
*Corresponding Author

Abstract
Climate misinformation often combines a factual proposition with a recognizable narrative, such as denying observed warming, rejecting human causation, minimizing impacts, attacking mitigation, or casting doubt on climate science. This paper presents a lightweight narrative-aware scientific claim verification agent for the official ClimateCheck setting. The revised evaluation uses the official annotated ClimateCheck training data, the official publications corpus of 394,269 abstracts, and a claim-level validation split of the annotated data. The public ClimateCheck test file is treated as a blind claim list because its public fields do not contain verification or narrative labels. The system combines hashed BM25, TF-IDF retrieval, latent semantic analysis, narrative-family probabilities, and a logistic-regression verifier. Full-corpus retrieval shows that BM25 remains the strongest first-stage retriever, with Recall@10 = 0.466, while the narrative-aware hybrid obtains Recall@10 = 0.444. In the judged candidate reranking setting, the narrative-aware ranker obtains the highest Candidate Recall@1 = 0.789 and MAP = 0.848, compared with 0.759 and 0.843 for TF-IDF. End-to-end verification remains difficult: the BM25 top-1 pipeline reaches Macro-F1 = 0.408, while the narrative-aware pipeline reaches Macro-F1 = 0.355. Claim-level narrative evaluation no longer produces a perfect score; single-label top-family Macro-F1 is 0.422, and fine-grained multi-label CARDS-code Macro-F1 is 0.098. These results show that narrative information is useful for reranking already plausible evidence candidates, but it does not replace strong lexical retrieval and does not by itself solve claim verification.

Keywords: *Scientific Claim Verification; Climate Misinformation; Evidence Ranking; Narrative Classification; Information Retrieva.l*

I. INTRODUCTION

Scientific claim verification requires a system to connect a short public claim with evidence from a longer scientific source and then decide whether the evidence supports, refutes, or leaves the claim unresolved. Climate communication makes this problem especially difficult because the claim is rarely only a factual proposition. It often carries a narrative frame: a cold-weather example may imply that warming has stopped, a statement about model uncertainty may imply that climate science is unreliable, and a statement about clean-energy costs may imply that mitigation is ineffective.

The ClimateCheck shared-task setting addresses this evidence-grounded problem by linking climate-related claims to scientific abstracts and asking systems to retrieve abstracts and classify the claim-abstract relation. The official task definition emphasizes retrieval from a large corpus of climate-related publications rather than only classification over preselected pairs. This distinction is important because a verifier cannot produce a reliable verdict if the evidence retriever does not first locate a relevant abstract.

Received on Januari 2026; Revised on Januari 2026; Accepted on March 2026; Published on April 2026

This revision narrows the paper's claim. The proposed agent is not presented as a new neural architecture or a replacement for dense retrieval. It is a transparent hybrid baseline that tests whether narrative-family probabilities can improve evidence ranking when combined with lexical matching. The main empirical question is therefore modest: does narrative information help rank evidence after the lexical retriever has already narrowed the evidence space, and does that improvement carry through to verification?

The revised study uses only official benchmark data for the ClimateCheck experiments. The annotated train split is divided by claim identifier into a model-training fold and a held-out validation fold. The official publications corpus is used for full-corpus abstract retrieval. Because the public test file does not expose gold verification or narrative labels, it is not used for supervised metric reporting. This design directly separates first-stage retrieval, judged candidate reranking, verification, narrative classification, and rationale alignment.

The results are intentionally interpreted cautiously. Narrative-aware reranking improves candidate evidence access, but BM25 remains the strongest first-stage full-corpus retriever, and the final verifier does not consistently benefit from narrative reranking. The contribution is therefore incremental and diagnostic: it clarifies where narrative signals help, where lexical matching remains stronger, and why stronger semantic retrieval and verification models are still needed.

II. LITERATURE REVIEW

Automated fact-checking developed from stance detection, textual entailment, and open-domain evidence retrieval. Early resources such as Emergent, the Fake News Challenge, and LIAR showed that claim-source classification benefits from lexical overlap, disagreement cues, and metadata, but many early benchmarks did not require systems to retrieve explicit evidence chains (Ferreira & Vlachos, 2016; Hanselowski et al., 2018; Wang, 2017). FEVER changed the standard by requiring systems to retrieve Wikipedia evidence and classify claims as supported, refuted, or not enough information (Thorne et al., 2018).

Scientific claim verification adds terminology, hedging, long sentences, and abstract-level context. SciFact demonstrated that scientific claims require evidence retrieval from paper abstracts and relation classification over domain-specific language (Wadden et al., 2020). Public-health verification also showed that evidence-centered explanations are important for user trust because a verdict without a rationale is difficult to inspect (Kotonya & Toni, 2020). Climate verification follows the same evidence-centered logic but involves climate-specific vocabulary and misinformation narratives (Diggelmann et al., 2020).

Retrieval is the first bottleneck in claim verification. BM25 remains a strong baseline because scientific claims and abstracts often share technical terms, named entities, measurements, and domain-specific phrases (Manning et al., 2008; Robertson & Zaragoza, 2009). Dense retrieval and late-interaction systems can improve semantic matching, but they require additional model resources and careful negative sampling (Karpukhin et al., 2020; Khattab & Zaharia, 2020). Sentence-BERT and related embedding methods are useful alternatives, although lexical models remain competitive in scientific domains where exact terminology carries evidence (Reimers & Gurevych, 2019).

Transformer classifiers such as BERT, RoBERTa, and SciBERT have improved natural language inference and scientific text classification (Beltagy et al., 2019; Devlin et al., 2019; Liu et al., 2019). However, a transformer classifier alone does not solve evidence retrieval. Retrieval-augmented systems also depend on the quality of the retrieved evidence before producing or explaining an answer (Lewis et al., 2020). For that reason, this paper evaluates retrieval and verification separately instead of reporting only claim-pair classification.

Narrative-aware verification connects fact-checking with misinformation analysis. A narrative describes the argumentative role of a claim, not only its topic. Two claims about Arctic sea ice may belong to different narratives if one discusses measured retreat and the other uses a short-term cold event to deny warming. Conversely, claims about different topics may share a narrative if both attack scientific institutions. The ClimateCheck data include CARDS-style narrative labels, which makes it possible to test whether narrative information helps evidence ranking. This paper treats narrative prediction as an auxiliary ranking signal rather than as an independent replacement for retrieval

III. RESEARCH METHOD

The experiment uses the official ClimateCheck annotated training file, the official ClimateCheck public test file, and the official ClimateCheck publications corpus. The annotated training file contains 3,023 claim-abstract pairs with verification labels and CARDS-style narrative labels. The public test file contains 176 rows, but the public fields for abstract, annotation, and narrative are empty; therefore, supervised metrics are reported on a claim-level validation split derived from the annotated training file. The publications corpus contains 394,269 abstracts and is used for full-corpus retrieval.

The train-validation split is made at the claim level with random seed 2025. This prevents the same claim from appearing in both model training and validation. The model-training fold contains 2,408 rows, 610 unique claims, and 1,941 unique abstracts. The validation fold contains 615 rows, 153 unique claims, and 584 unique abstracts. Table 1 lists the fields used by the

pipeline, and Table 2 reports the split profile. Figure 1 and Figure 2 show the verification-label and narrative-family distributions in the official annotated data.

Table 1. Dataset fields and their roles in the experimental pipeline.

Field	Role	Used by
claim	Query text	Retrieval, verification, narrative classification
abstract	Candidate scientific abstract	Reranking, verification, evidence sentence selection
abstract_id	Evidence identifier	Retrieval and ranking metrics
claim_id	Claim identifier	Claim-level split and grouping
annotation	Supports, Refutes, or Not Enough Information	Verification and evidentiary-reranking labels
narrative	CARDS-style narrative code	Narrative mapping and classifier training
publications corpus	Large retrieval index	Full-corpus abstract retrieval
SciFact rationales	Gold evidence sentences in auxiliary data	Rationale-alignment analysis

The pipeline has five components, shown in Figure 3. First, a full-corpus retriever ranks abstracts from the publications corpus. Second, a candidate reranker compares judged claim-abstract candidates using lexical and narrative signals. Third, a verification classifier predicts Supports, Refutes, or Not Enough Information for the selected claim-abstract pair. Fourth, narrative classifiers predict either the top-level family or the fine-grained CARDS code. Fifth, a sentence-selection module is evaluated on SciFact, because the ClimateCheck fields used here do not include a gold rationale sentence.

Table 2. Dataset and split profile used in the revised evaluation.

Split/source	Rows	Unique claims	Unique abstracts	Supports	Refutes	Not Enough Information
Official ClimateCheck annotated train	3023	763	2382	1399	451	1173
Model-training fold	2408	610	1941	1100	370	938
Claim-level validation fold	615	153	584	299	81	235
Public ClimateCheck test claims	176	172	0	0	0	0
Publications corpus	394269	0	394269	0	0	0

The narrative-aware score combines normalized BM25, TF-IDF, and narrative compatibility. Narrative compatibility is the dot product between the claim narrative probability vector and the abstract narrative probability vector. The candidate-reranking weights are selected by grid search on an inner development split of the model-training fold. The best development setting assigns 0.40 to BM25, 0.40 to TF-IDF, 0.00 to character n-grams, and 0.20 to narrative compatibility. For full-corpus retrieval, the same non-character components are used with the same proportions. This explains the weighting scheme rather than treating it as a fixed heuristic.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 327-340
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

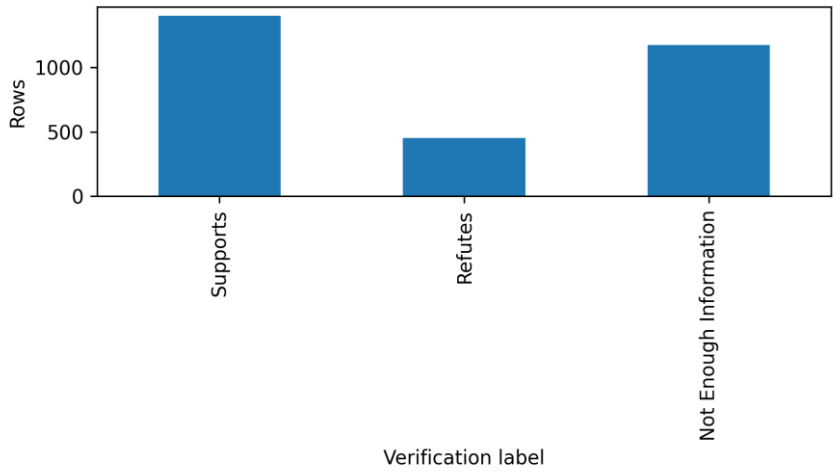

Figure 1. Verification-label distribution in the official annotated ClimateCheck training data.

Full-corpus retrieval is evaluated using evidentiary abstracts, defined as validation pairs labeled Supports or Refutes. Not Enough Information pairs are kept for verification but are not counted as evidentiary targets in retrieval. Candidate reranking uses the judged validation candidates for each claim and asks whether evidentiary abstracts are ranked above Not Enough Information candidates. This distinction keeps the full-corpus retrieval task separate from the smaller reranking task.

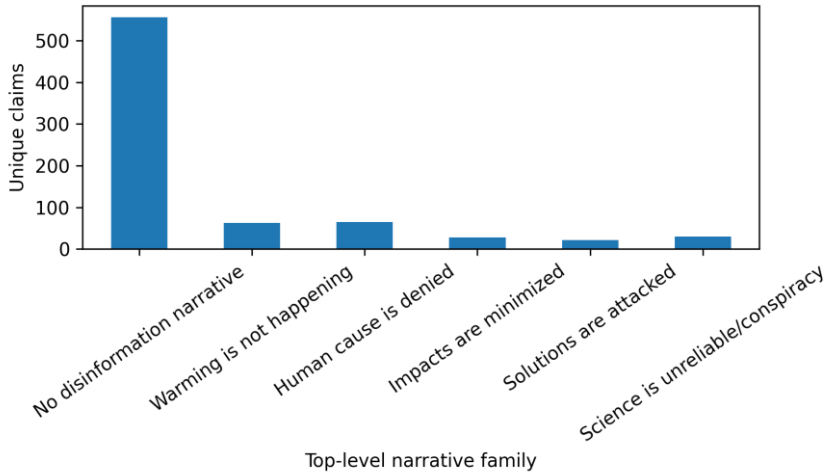

Figure 2. Top-level narrative-family distribution over unique claims in the official annotated data.

The verification model is trained on claim-abstract pairs from the model-training fold. It receives the claim and selected abstract as a single text pair and predicts one of the three ClimateCheck labels. The reported end-to-end verification metric uses the top-ranked candidate selected by each

candidate ranker. This avoids reporting only gold-pair classification, which would hide retrieval and reranking errors.

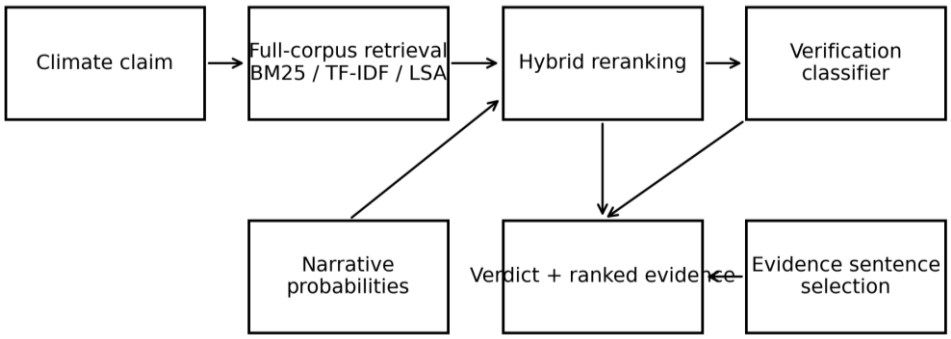

Figure 3. Architecture of the revised narrative-aware verification pipeline.

Narrative classification is evaluated at the claim level. The revised evaluation reports both single-label top-family prediction and multi-label prediction over CARDS-style codes. This stricter setup directly addresses the earlier risk of reporting a perfect narrative-classification score from an easier top-level or lexically trivial split.

Table 3. Model configurations used in the revised evaluation.

Model	Main configuration	Evaluation role
Hashed BM25	BM25 scoring over a fixed 32,768-dimensional unigram hashing vocabulary	Full-corpus retrieval
TF-IDF cosine	Hashed unigram counts with TF-IDF normalization	Full-corpus retrieval
LSA semantic retrieval	48-dimensional truncated SVD over the TF-IDF index	Semantic retrieval baseline
SciLex lexical fusion	0.60 word TF-IDF + 0.40 character 3-5 gram cosine	Candidate reranking
Narrative-aware agent	0.40 BM25 + 0.40 TF-IDF + 0.20 narrative compatibility	Weights tuned on inner development claims
Verification agent	Claim-abstract TF-IDF logistic regression	Balanced logistic regression
Narrative classifiers	Claim TF-IDF logistic regression and one-vs-rest models	Top-family and CARDS-code prediction

For rationale alignment, the paper uses SciFact as an auxiliary scientific fact-checking dataset because its claims include sentence-level evidence annotations. Four sentence selectors are compared: first sentence, lexical overlap, TF-IDF sentence ranking, and an evidence keyword ranker that combines TF-IDF similarity, lexical overlap, and a simple evidence-position prior.

IV. RESULT AND DISCUSSION

Table 4 and Figure 4 present the full-corpus retrieval results over the 394,269-abstract publications corpus. BM25 is the strongest first-stage model, reaching Recall@10 = 0.466 and nDCG@10 = 0.190. The narrative-aware hybrid obtains Recall@10 = 0.444 and nDCG@10 = 0.183. This result is important because it prevents the paper from overstating the narrative signal:

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 327-340
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

narrative compatibility is helpful in the reranking setting, but it does not outperform BM25 as a first-stage retriever over the full corpus.

Table 4. Full-corpus abstract retrieval results on the claim-level validation fold.

Model	Recall@2	Recall@5	Recall@10	MRR	nDCG@10	First rank
Random baseline	0.000	0.000	0.000	0.000	0.000	121196.278
BM25	0.248	0.368	0.466	0.278	0.190	635.301
TF-IDF cosine	0.128	0.248	0.368	0.173	0.120	689.759
LSA semantic retrieval	0.015	0.038	0.053	0.025	0.010	6375.113
Narrative-aware agent	0.241	0.346	0.444	0.264	0.183	547.331

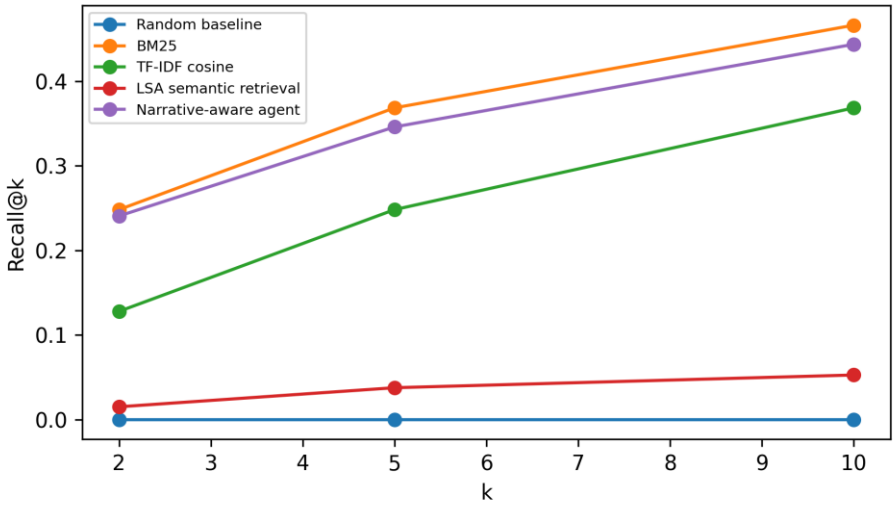

Figure 4. Full-corpus Recall@k curves for retrieval models.

The LSA semantic baseline performs worse than BM25 and TF-IDF in this configuration. This does not rule out stronger dense retrievers or transformer rerankers, but it shows that a simple low-dimensional semantic model is not sufficient for this benchmark. The result is consistent with the importance of exact scientific terminology in climate abstracts.

Table 5. Candidate evidence reranking results on judged validation candidates.

Model	Recall@1	Recall@3	Recall@5	MAP	MRR	nDCG@5
Random baseline	0.692	0.962	1.000	0.796	0.816	0.860
BM25	0.744	0.970	1.000	0.822	0.861	0.883
TF-IDF cosine	0.759	0.970	1.000	0.843	0.863	0.893
SciLex lexical fusion	0.737	0.940	1.000	0.823	0.848	0.880
Narrative-aware agent	0.789	0.962	1.000	0.848	0.876	0.900

Table 5 and Figure 5 show the judged candidate reranking results. In this smaller setting, where candidate abstracts are already associated with a claim, the narrative-aware agent achieves the best Candidate Recall@1 = 0.789 and MAP = 0.848. The gain over TF-IDF is modest: Recall@1 rises from 0.759 to 0.789, and MAP rises from 0.843 to 0.848. This supports a cautious

interpretation: narrative information helps reorder plausible evidence candidates, but the improvement is incremental rather than transformative.

Table 6. Top-1 candidate reranking plus verification results.

Pipeline	Top-1 evid.	Accuracy	Macro-F1	Weighted-F1
BM25 + verification agent	0.647	0.464	0.408	0.462
TF-IDF cosine + verification agent	0.660	0.399	0.313	0.397
SciLex lexical fusion + verification agent	0.641	0.425	0.351	0.423
Narrative-aware agent + verification agent	0.686	0.444	0.355	0.447

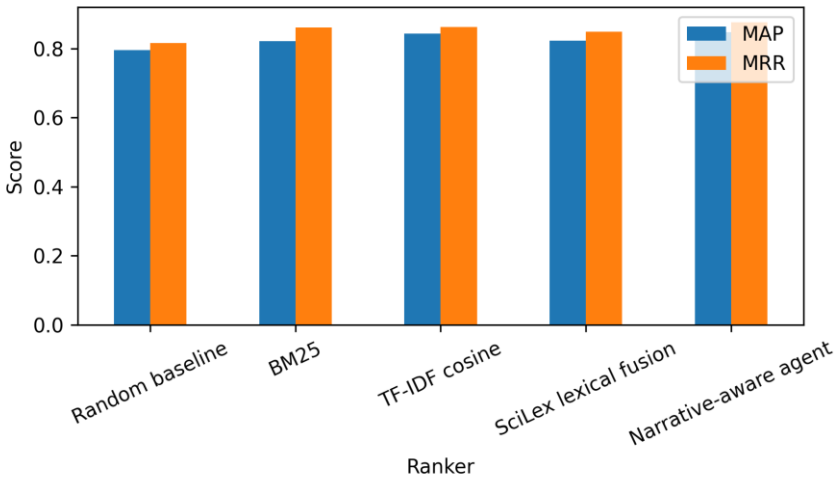

Figure 5. Candidate reranking MAP and MRR by ranker.

Table 6 and Figure 6 report end-to-end top-1 verification after candidate selection. The BM25 pipeline obtains the strongest Macro-F1 = 0.408, while the narrative-aware pipeline obtains Macro-F1 = 0.355. The narrative-aware pipeline has the highest top-1 evidentiary rate, but that does not translate into the best final label F1. This shows that retrieval and verification errors are related but not identical. A retrieved evidentiary abstract may still be difficult for a lightweight classifier, especially when the claim and abstract differ in wording or when the correct distinction is between Refutes and Not Enough Information.

Table 7. Narrative classification results under claim-level validation.

Model	Task	Accuracy/ exact	Macro-F1	Micro-F1
Majority top-family prior	single-label top family	0.752	0.143	0.752
Claim TF-IDF logistic regression	single-label top family	0.686	0.422	0.686
Claim TF-IDF one-vs-rest	multi-label top family	0.647	0.355	0.705
Claim TF-IDF one-vs-rest	multi-label CARDS code	0.601	0.098	0.693

Table 7 and Figure 7 replace the earlier perfect narrative-classification result with a stricter claim-level evaluation. The majority top-family prior reaches high accuracy because no-disinformation claims dominate the data, but its Macro-F1 is only 0.143. The claim TF-IDF logistic-regression model reaches top-family Macro-F1 = 0.422. The multi-label CARDS-code task is much harder,

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 327-340
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

with Macro-F1 = 0.098. These results make the narrative component more credible: the task is not reported as solved, and the fine-grained codes are clearly more difficult than top-level families.

Table 8. Candidate reranking ablation results for the narrative-aware agent.

Model	Recall@1	Recall@3	Recall@5	MAP	MRR	nDCG @5
Narrative-aware agent	0.789	0.962	1.000	0.848	0.876	0.900
Agent - narrative score	0.782	0.955	1.000	0.843	0.873	0.896
Agent - char n-grams	0.789	0.962	1.000	0.848	0.876	0.900
Agent - BM25	0.737	0.955	1.000	0.833	0.848	0.886
Agent - TF-IDF	0.729	0.962	1.000	0.815	0.850	0.878

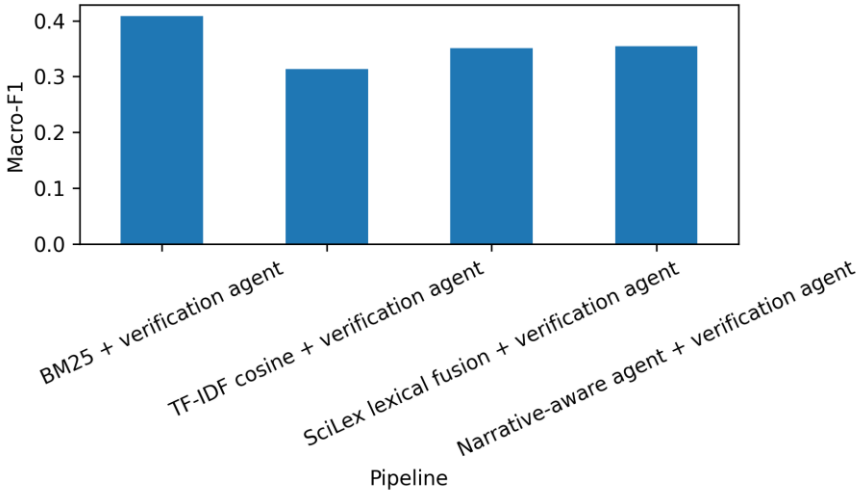

Figure 6. End-to-end verification Macro-F1 by candidate-selection pipeline.

Table 8 and Figure 8 report the candidate-reranking ablation. Removing the narrative score lowers Recall@1 from 0.789 to 0.782 and MAP from 0.848 to 0.843. Removing BM25 or TF-IDF causes a larger decline. Character n-grams receive zero weight in the tuned setting and therefore do not change the score in this split. The ablation reinforces the main interpretation: narrative compatibility provides a small reranking gain, while lexical matching remains the dominant evidence signal.

Table 9. Error analysis by top-level narrative family.

Narrative family	Claims	Cand. R@1	Macro-F1	Top-family F1
No disinformation narrative	115	0.800	0.308	0.144
Warming is not happening	13	0.750	0.303	0.233
Human cause is denied	7	0.857	0.467	1.000
Impacts are minimized	7	0.833	0.167	0.083
Solutions are attacked	6	0.750	0.095	0.167
Science is unreliable/conspiracy	5	0.500	0.222	0.111

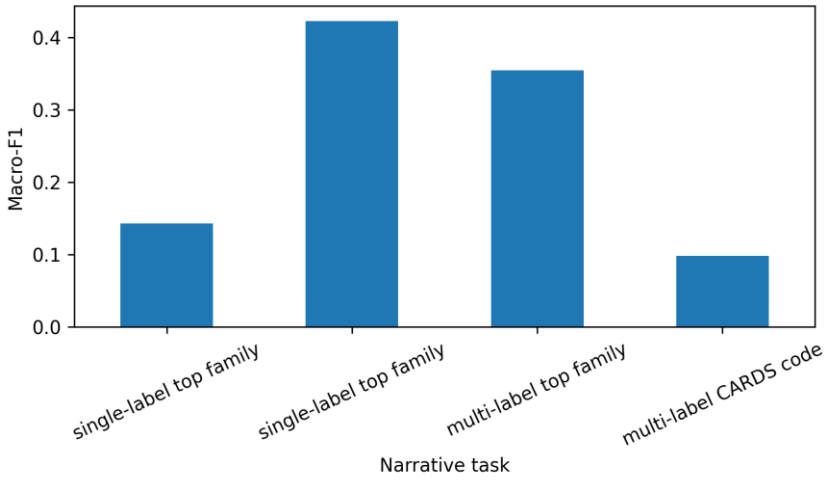

Figure 7. Narrative-classification Macro-F1 under stricter claim-level evaluation.

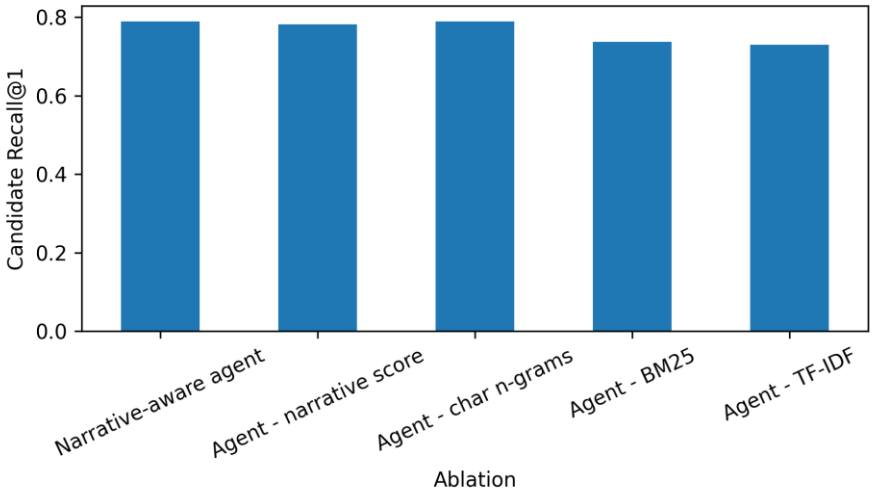

Figure 8. Candidate Recall@1 under signal ablations.

Table 9 reports error patterns by top-level narrative family. Most validation claims belong to the no-disinformation family, so its results are the most stable. The smaller families contain between five and thirteen validation claims, so one or two errors can change the family-level score substantially. Human-causation denial obtains the strongest top-family F1 in this split, while other minority families remain difficult. The main diagnostic result is that family-level narrative prediction is uneven and should not be treated as a solved task.

Table 10. Auxiliary SciFact rationale-alignment results.

Selector	Exact match	Token-F1
First sentence	0.057	0.218
Lexical overlap	0.455	0.465
TF-IDF sentence ranker	0.502	0.509
Evidence keyword ranker	0.517	0.515

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 327-340
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Table 10 and Figure 9 present the auxiliary SciFact rationale-alignment results. Because the ClimateCheck fields used in the main experiment do not contain a gold rationale sentence, this auxiliary experiment is used only to test sentence-selection behavior in a scientific fact-checking setting. The evidence keyword ranker reaches Exact Match = 0.517 and Token-F1 = 0.515, slightly above the TF-IDF sentence ranker. This result supports the use of extractive sentence selection, but it should not be interpreted as a ClimateCheck rationale score.

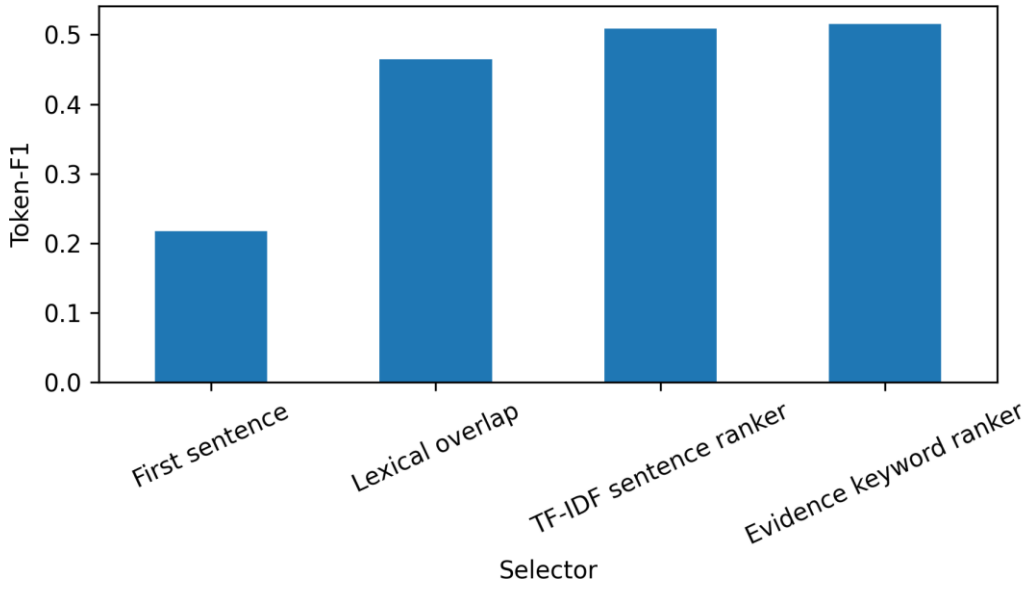

Figure 9. Auxiliary SciFact rationale Token-F1 by sentence selector.

V. CONCLUSION AND RECOMMENDATION

This paper presented a revised narrative-aware scientific claim verification agent for the official ClimateCheck setting. The revision clarifies the dataset provenance, separates the public unlabeled test file from the supervised validation evaluation, adds full-corpus retrieval over the official publications corpus, introduces a semantic LSA baseline, tunes the ranker weights on an inner development split, and replaces the earlier perfect narrative-classification result with stricter claim-level metrics.

The empirical conclusion is more cautious than the original version. Narrative information helps candidate reranking, increasing Candidate Recall@1 from 0.759 under TF-IDF to 0.789 under the narrative-aware agent. However, BM25 remains the strongest full-corpus first-stage retriever, and the narrative-aware pipeline does not produce the best end-to-end verification Macro-F1. Therefore, the main contribution is not a claim of large performance improvement. It is a transparent analysis showing where narrative information helps and where lexical retrieval and stronger verification models remain necessary.

The most important limitation is first-stage retrieval from the full 394,269-abstract corpus. The best full-corpus Recall@10 is 0.466, which leaves many evidentiary abstracts outside the first ten results. Future work should add stronger dense retrieval, hard-negative training from same-topic abstracts, and transformer cross-encoder reranking. A second limitation is verification. The lightweight logistic-regression verifier is useful as a transparent baseline, but it struggles with fine distinctions among Supports, Refutes, and Not Enough Information. A scientific transformer verifier with calibrated uncertainty would be a natural next step.

The stricter narrative results also show that top-level narrative families are easier than fine-grained CARDS codes. Future narrative-aware systems should evaluate multi-label codes and should avoid relying only on broad narrative families. For deployment, a practical climate-claim verification agent should combine strong lexical retrieval, a semantic reranker, narrative-aware candidate ordering, and extractive evidence sentences, with human review of the ranked evidence before accepting a final verdict.

REFERENCES

Abu Ahmad, R., Upravitelev, M., Usmanova, A., Solopova, V., & Rehm, G. (2025). The ClimateCheck shared task: Scientific fact-checking of social media claims about climate change. Proceedings of the 5th Workshop on Scholarly Document Processing.

Abu Ahmad, R., Upravitelev, M., Usmanova, A., Solopova, V., & Rehm, G. (2026). ClimateCheck 2026: Scientific fact-checking and disinformation narrative classification of climate-related claims. arXiv preprint arXiv:2603.26449.

Augenstein, I., Lioma, C., Wang, D., Chaves Lima, L., Hansen, C., Hansen, C., & Simonsen, J. G. (2019). MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. Proceedings of EMNLP-IJCNLP, 4685-4697.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. Proceedings of EMNLP-IJCNLP, 3615-3620.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. Proceedings of EMNLP, 632-642.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). S2ORC: The Semantic Scholar Open Research Corpus. Proceedings of ACL, 4969-4983.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL, 4171-4186.

Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., & Leippold, M. (2020). CLIMATE-FEVER: A dataset for verification of real-world climate claims. arXiv preprint arXiv:2012.00614.

Ferreira, W., & Vlachos, A. (2016). Emergent: A novel data-set for stance classification. Proceedings of NAACL, 1163-1168.

Hanselowski, A., PVS, A., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M., & Gurevych, I. (2018). A retrospective analysis of the Fake News Challenge stance-detection task. Proceedings of COLING, 1859-1874.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. Proceedings of EMNLP, 6769-6781.

Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. Proceedings of SIGIR, 39-48.

Kotonya, N., & Toni, F. (2020). Explainable automated fact-checking for public health claims. Proceedings of EMNLP, 7740-7754.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Lewis, M., Yih, W. T., Rocktaschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Manning, C. D., Raghavan, P., & Schutze, H. (2008). Introduction to information retrieval. Cambridge University Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv preprint arXiv:2205.01833.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of EMNLP-IJCNLP, 3982-3992.

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4), 333-389.

Schuster, T., Fisch, A., & Barzilay, R. (2021). Get your vitamin C! Robust fact verification with contrastive evidence. Proceedings of NAACL, 624-643.

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. Proceedings of NAACL, 809-819.

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., & Hajishirzi, H. (2020). Fact or fiction: Verifying scientific claims. Proceedings of EMNLP, 7534-7550.

Wang, W. Y. (2017). Liar, liar pants on fire: A new benchmark dataset for fake news detection. Proceedings of ACL, 422-426.

Williams, A., Nangia, N., & Bowman, S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. Proceedings of NAACL, 1112-1122.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, S., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Xu, C., Le Scao, T., Gugger, S., & Rush, A. (2020). Transformers: State-of-the-art natural language processing. Proceedings of EMNLP: System Demonstrations, 38-45.