AlexUNLP-FMT at ClimateCheck Shared Task: Hybrid Retrieval with Adaptive Similarity Graph-based Reranking for Climate-related Social Media Claims Fact Checking

Mahmoud Fathallah, Nagwa ElMakky, Marwan Torki
Department of Computer and Systems Engineering
Alexandria University, Egypt
{es-mahmodfath96, nagwamakky, mtorki}@alexu.edu.eg

Abstract

In this paper, we describe our work for the ClimateCheck shared task at the Scholarly Document Processing (SDP) workshop, ACL 2025. We focus on Subtask 1: Abstracts retrieval. The task involves retrieving relevant abstracts from a large corpus to verify claims made on social media about climate change. We explore various retrieval and reranking techniques, including fine-tuning transformer-based dense retrievers, sparse retrieval methods, and reranking using cross-encoder models. Our final and best-performing system utilizes a hybrid retrieval approach combining BM25 sparse retrieval with a fine-tuned Stella model for dense retrieval, followed by an MSMARCO-trained MiniLM cross-encoder model for reranking. We adapt an iterative graph-based reranking approach that leverages a document similarity graph built over the document corpus to update the candidate pool for reranking dynamically. Our system achieved a score of 0.415 on the final test set for Subtask 1, securing third place on the final leaderboard.

Our code is available on GitHub[1].

1 Introduction

Misinformation spreading on social media poses a significant threat to public understanding of scientific issues, particularly in domains such as climate change, where accurate information is needed to raise awareness and create evidence-based policies.

Social media platforms are often the first point of exposure to climate-related content for the general public, making it easy for misleading claims and information to spread. Therefore, there is a need for automated fact-checking systems that can assess the veracity of such claims in real time.

Automated evidence-based fact-checking remains a challenging task (Glockner et al., 2022). Highly effective retrieval module that can retrieve relevant evidence to support or refute a given claim is a necessary component of the evidence-based fact-checking system (Zheng et al., 2024).

This paper presents our approach for Subtask 1 of the ClimateCheck shared task (Abu Ahmad et al., 2025b), held at the Scholarly Document Processing (SDP) workshop at ACL 2025. The Subtask focuses on retrieving relevant scientific abstracts from a large corpus in response to climate-related claims made on social media.

We experiment with dense and sparse retrieval models and employ a retrieval-reranking pipeline. We fine-tune models using supervised contrastive learning and evaluate the effectiveness of hybrid retrieval pipelines that combine sparse and dense approaches. We adapt a graph-based reranking approach inspired by prior work on corpus graph expansion (MacAvaney et al., 2022), where the reranking pool is iteratively enriched using neighbors of top-ranked documents.

2 Related Work

Information retrieval (IR) pipelines generally rely on sparse or dense retrieval techniques.

Sparse retrieval Sparse retrieval methods, such as BM25 (Robertson and Zaragoza, 2009), represent queries and documents as high-dimensional sparse vectors based on term frequency-inverse document frequency statistics (TF-IDF). While effective in capturing lexical similarity, these models often struggle to capture semantic similarity.

Dense Retrieval Dense retrieval models (Karpukhin et al., 2020; Xiao et al., 2023; Zhang et al., 2025) address the problem of semantic similarity faced by sparse retrieval models. They embed both queries and documents to obtain dense vector representations that allow similarity-based search through vector similarity.

Reranking Reranking is a critical step in IR

[1] https://github.com/Mahmoud-Mohammed-Fathallah/climatecheck-shared-task

Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 288–292

Label	Training
Supports	446
Refutes	241
Not enough info.	457
Total	1144

Table 1: Distribution of labels in the provided training set.

Label	Training	Validation
Supports	360	86
Refutes	196	45
Not enough info.	361	96
Total	917	227

Table 2: Distribution of labels in our training and validation sets.

pipelines (Liu et al., 2025). Bi-encoder models enable efficient and fast retrieval but require reranking to enhance performance by utilizing a cross-encoder model to jointly encode query-document pairs and output a similarity score (Nogueira and Cho, 2020). This allows for deeper interaction between queries and documents, further enhancing the performance.

Adaptive reranking Adaptive reranking techniques that utilize similarity graphs (MacAvaney et al., 2022; Rathee et al., 2025) have been developed to overcome the limitations of standard retrieval-reranking pipelines, where the reranking performance is limited by the set initially retrieved by the retriever (MacAvaney et al., 2022). In the adaptive reranking approach, a similarity graph is used to retrieve documents related to the top-ranked ones, enabling richer reranking candidates.

3 Data

The shared task provided a training set, a test set, and a document corpus (Abu Ahmad et al., 2025a). The training set included 1,144 claim–abstract pairs labeled as supports, refutes, or not enough information; the label distribution is shown in Table 1. The retrieval corpus contained 394,269 paper abstracts. The test set consisted of 176 unlabeled claims.

Since the shared task initially provided only a training set, we created our own validation set by randomly sampling 50 unique claims and their associated data points from the original training set. The remaining samples formed our training set. Distributions for both sets are shown in Table 2.

4 Methodology

In this section, we outline our approach and the final submitted system.

4.1 Retrieval models

We first explored different bi-encoder models for dense retrieval, such as bge-large-en[2] (Xiao et al., 2023), stella-en-400M-v5[3] (Zhang et al., 2025), and inf-retriever-v1-1.5b[4] (Junhan Yang, 2025). We fine-tuned the first two using contrastive learning (Qiu et al., 2021) utilizing Multiple Negatives Ranking Loss to bring embeddings of related query-abstract pairs closer together and separate unrelated ones. We also experimented with BM25 (Robertson and Zaragoza, 2009), a traditional lexical search algorithm used as a strong baseline and in hybrid approaches.

4.2 Reranking models

We experimented with different cross-encoder models for reranking, comparing the powerful fine-tuned model bge-reranker-v2-m3[5] (Chen et al., 2024) to other models trained on the MS-MARCO dataset (Nguyen et al., 2016), such as ms-marco-MiniLM-L12-v2[6] (Wang et al., 2020), ms-marco-electra-base[7] (Clark et al., 2020), and reranker-msmarco-ModernBERT-base-lambdaloss[8] (Warner et al., 2024).

4.3 Hybrid retrieval

To enhance retrieval performance, we integrated dense retrieval with sparse retrieval, leveraging the strong lexical matching capabilities of methods like BM25(Robertson and Zaragoza, 2009) alongside the powerful semantic search capabilities of dense models such as Stella bi-encoder model (Zhang et al., 2025) to enhance retrieval performance. The top-k documents from both models were combined, with duplicates removed.

[2] https://huggingface.co/BAAI/bge-large-en

[3] https://huggingface.co/NovaSearch/stella_en_400M_v5

[4] https://huggingface.co/infly/inf-retriever-v1-1.5b

[5] https://huggingface.co/BAAI/bge-reranker-v2-m3

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

[7] https://huggingface.co/cross-encoder/ms-marco-electra-base

[8] https://huggingface.co/tomaarsen/reranker-msmarco-ModernBERT-base-lambdaloss

4.4 Similarity graph-based reranking

To address the limitations of the initial retrieval stage, where highly relevant documents may be missing from the retrieved set, we implemented an adaptive retrieval and reranking strategy.

We first constructed a similarity graph over the entire corpus, connecting each document to its top-k semantically similar neighbors. The adaptive reranking strategy proceeds as follows: an initial set of documents is retrieved using a retriever, and the top-n candidates are ranked by a cross-encoder reranker. The top-p documents (p < n) are selected and expanded by including their neighbors from the similarity graph. This augmented set is reranked, and the process is repeated for a fixed number of iterations.

4.5 Final system

Our final system used a graph built with all-MiniLM-L6-v2 bi-encoder model[9] with k=10 nearest neighbors per document and n=20 iterations for the adaptive reranking step. These values were selected after experimenting with different parameters to balance retrieval performance with computational efficiency, allowing the system to explore a broader set of relevant documents through multiple reranking iterations while keeping the runtime feasible. For reranking, we used "ms-marco-MiniLM-L12-v2" cross-encoder model. The initial retrieval stage combined the top 50 documents retrieved by BM25 sparse retrieval and the fine-tuned stella-en-400M-v5 dense retriever. The value 50 is chosen as a reasonable default, as we were unable to perform extensive hyperparameter tuning due to time limitations. The full system architecture is illustrated in Figure 1.

5 Experiments and Results

Evaluation on the validation and test sets was performed using the official shared task metrics: Recall@k (R@k for k=2, 5, 10), B-pref, and the overall SubtaskI-Score, defined as the average of the other metrics.

5.1 Training details

All experiments were conducted on a single NVIDIA V100 GPU. During training, we used batch sizes of 8 and 16, and optimized the models using the Adam optimizer with learning rates of 1e-5, 1e-6, and 4e-6.

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

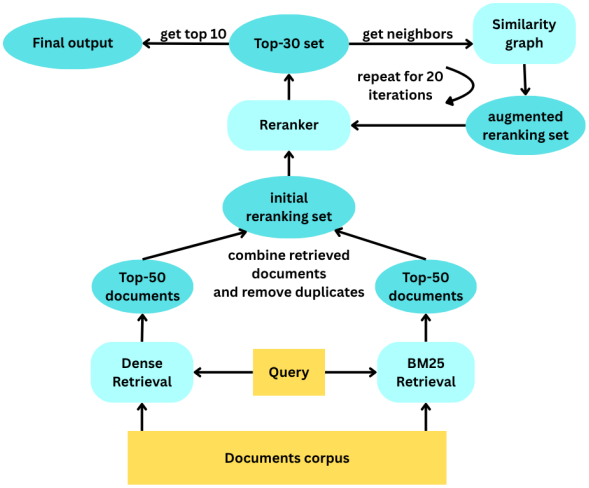

Figure 1: Our final system that utilizes hybrid retrieval with adaptive reranking.

5.2 Retrieval models

We compared BM25 (Robertson and Zaragoza, 2009), bge-large-en (Xiao et al., 2023), stella-en-400M-v5 (Zhang et al., 2025) and inf-retriever-v1-1.5b (Junhan Yang, 2025), both with and without fine-tuning on the training set using Multiple Negatives Ranking Loss. The fine-tuned stella model outperformed the other retrieval models. Results on the validation set are shown in Table 3.

5.3 Reranking models

To choose a reranking model, we compared several cross-encoder models: bge-reranker-v2-m3 (Chen et al., 2024), ms-marco-MiniLM-L12-v2 (Wang et al., 2020), ms-marco-electra-base (Clark et al., 2020), and reranker-msmarco-ModernBERT-base-lambdaloss (Warner et al., 2024). The bge-reranker model was fine-tuned on the training set using Multiple Negatives Ranking Loss, while the other models, trained on MS-MARCO dataset (Nguyen et al., 2016), were used without fine-tuning. For a fair comparison, we fixed BM25 as the initial retriever and applied each reranker to the same retrieved set. Results showed that the ms-marco-MiniLM-L12-v2 model outperformed the other models. Validation set results are presented in Table 4.

5.4 Hybrid retrieval

For the hybrid retrieval experiment, we combined the top 50 retrieved documents from BM25 and our best dense retrieval model, the fine-tuned stella-400M. We then rerank this initial set using our best

Model	R@2	R@5	R@10	B-pref	score
BM25	0.0977	0.1407	0.2051	0.1363	0.1450
bge-large	0.0255	0.1122	0.1422	0.1515	0.1078
bge-large*	**0.1351**	0.1840	0.2414	0.2112	0.1930
inf-retriever	0.0633	0.1774	0.2670	0.2676	0.1938
inf-retriever*	0.0785	**0.2044**	0.2862	0.2441	0.2033
stella*	0.1218	0.1851	**0.2911**	**0.2777**	**0.2189**

Table 3: Comparing different retrieval models on the validation set. The * in the model name means that it is fine-tuned on the training set. The best results are in bold.

Model	R@2	R@5	R@10	B-pref	Score
bge-reranker*	0.1440	0.2722	0.3570	0.2840	0.2643
ModernBERT	0.1807	0.2981	0.3918	0.3249	0.2989
electra-base	0.1381	0.3003	0.3644	0.2954	0.2746
MiniLM-L12	**0.1918**	**0.4144**	**0.6281**	**0.4053**	**0.4099**

Table 4: Comparing different Reranking models on the validation set. The * in the model name means that it is fine-tuned on the training set. The best results are in bold.

Metric	BM25	Stella*	Hybrid
R@2	0.1918	0.2303	**0.2344**
R@5	**0.4144**	0.4059	0.3933
R@10	0.6281	0.5640	**0.6466**
B-pref	0.4053	0.4259	**0.4297**
score	0.4099	0.4065	**0.4260**

Table 5: Results of Hybrid retrieval (BM25 + fine-tuned Stella) compared to each model alone with reranking using MiniLM-L12 on the validation set.

Metric	Validation	Test
R@2	0.2225	0.2099
R@5	0.4155	0.3962
R@10	0.6533	0.5911
B-pref	0.5162	0.4634
Score	0.4519	0.4152

Table 6: Results of our final system on the validation and test sets.

reranking model, ms-marco-MiniLM-L12-v2. To demonstrate the value of hybrid retrieval, we compared its results to those of each individual model. Results show that hybrid retrieval outperforms both models. Validation set results are shown in Table 5.

5.5 Graph-based adaptive reranking

For the final submission, we used the graph-based adaptive reranking approach, as illustrated in section 4.5. Results on the validation set showed that this adaptive reranking method improved the overall score by approximately 2.6% compared to the hybrid retrieval approach alone. The system's results on both the validation and test sets are shown in Table 6.

6 Conclusion

In this paper, we presented a hybrid retrieval system with adaptive reranking for evidence retrieval for climate-related social media claims, developed for Subtask 1 of the ClimateCheck shared task. Our system combined BM25-based sparse retrieval with a fine-tuned dense retriever, followed by a graph-based adaptive reranking approach utilizing a document similarity graph. We demonstrated that hybrid retrieval paired with iterative reranking significantly improved retrieval effectiveness, achieving third place in the final leaderboard.

Our findings emphasize the importance of combining hybrid retrieval with adaptive reranking to enhance the performance of scientific evidence retrieval systems. The use of graph-based expansion enabled the discovery of relevant abstracts that were missed by standard top-k methods.

Limitations

Despite the competitive performance of our adaptive reranking approach, several limitations remain. We did not explore careful tuning of hyperparameters, such as the number of neighbors in the similarity graph or the number of reranking iterations. Additionally, we did not explore the use of different models for constructing the similarity graph.

References

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

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

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

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *Preprint*, arXiv:2003.10555.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders nlp fact-checking unrealistic for misinformation. *Preprint*, arXiv:2210.13865.

Yichen Yao Wei Chu Yinghui Xu Yuan Qi Junhan Yang, Jiahe Wan. 2025. inf-retriever-v1 (revision 5f469d7).

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.

Qi Liu, Haozhe Duan, Yiqun Chen, Quanfeng Lu, Weiwei Sun, and Jiaxin Mao. 2025. Llm4ranking: An easy-to-use framework of utilizing large language models for document reranking. *Preprint*, arXiv:2504.07439.

Sean MacAvaney, Nicola Tonellotto, and Craig Macdonald. 2022. Adaptive re-ranking with a corpus graph. In *Proceedings of the 31st ACM International Conference on Information amp; Knowledge Management*, CIKM '22, page 1491–1500. ACM.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage re-ranking with bert. *Preprint*, arXiv:1901.04085.

Hefei Qiu, Wei Ding, and Ping Chen. 2021. Contrastive learning of sentence representations. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 277–283, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Mandeep Rathee, Sean MacAvaney, and Avishek Anand. 2025. Quam: Adaptive retrieval through lt;ugt;qult;/ugt;ery lt;ugt;alt;/ugt;ffinity lt;ugt;mlt;/ugt;odelling. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, WSDM '25, page 954–962. ACM.

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

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2025. Jasper and stella: distillation of sota embedding models. *Preprint*, arXiv:2412.19048.

Liwen Zheng, Chaozhuo Li, Xi Zhang, Yu-Ming Shang, Feiran Huang, and Haoran Jia. 2024. Evidence retrieval is almost all you need for fact verification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9274–9281, Bangkok, Thailand. Association for Computational Linguistics.