ClimateCheck2025: Multi-Stage Retrieval Meets LLMs for Automated Scientific Fact-Checking

Anna Kiepura[†], **Jessica Lam**[†]
[†]Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland
{akiepura, lamjessica}@ini.ethz.ch

Abstract

Misinformation on social media poses significant risks, particularly when it concerns critical scientific issues such as climate change. One promising direction for mitigation is the development of automated fact-checking systems that verify claims against authoritative scientific sources. In this work, we present our solution[1] to the ClimateCheck2025 shared task, which involves retrieving and classifying scientific abstracts as evidence for or against given claims. Our system is built around a multi-stage hybrid retrieval pipeline that integrates lexical, sparse neural, and dense neural retrievers, followed by cross-encoder and large language model (LLM)-based reranking stages. For stance classification, we employ prompting strategies with LLMs to determine whether a retrieved abstract supports, refutes, or provides no evidence for a given claim. Our approach achieves the second-highest overall score across both subtasks of the benchmark and significantly surpasses the final baseline by 53.76% on Subtask I score (defined as an average across Recall@2, Recall@5, Recall@10, and B-Pref). Notably, we achieve state-of-the-art performance in Recall@2. These results highlight the effectiveness of combining structured retrieval architectures with the emergent reasoning capabilities of LLMs for scientific fact verification, especially in domains where reliable human annotation is scarce and timely intervention is essential.

1 Introduction

The rapid proliferation of online misinformation, particularly in scientific, health, and policy contexts, has intensified the demand for reliable automated fact-checking systems (Li and Chang, 2022; Schlicht et al., 2023). These systems aim to assess the veracity of natural language claims by retrieving and evaluating relevant evidence from large text corpora. This process hinges on two core challenges: (**1**) retrieving relevant information from vast knowledge sources, and (**2**) determining whether the retrieved content supports, refutes, or fails to inform the claim.

Traditional keyword-based retrieval methods often struggle with these tasks, especially in domains requiring deep semantic understanding or domain-specific reasoning (Urbani et al., 2024; Devasier et al., 2025). Recent advances in neural retrievers and large language models (LLMs) have improved retrieval and reasoning capabilities across diverse topics (Vykopal et al., 2024; Quelle and Bovet, 2024; Ou et al., 2025). Nonetheless, integrating high-recall retrieval with robust, claim-sensitive reasoning remains a key bottleneck - particularly in scientific domains, where evidence is often sparse, nuanced, and hedged (Hyland, 1996).

In this paper, we present our system for the ClimateCheck2025 shared task (Abu Ahmad et al., 2025b), which consists of two subtasks: (**1**) for each climate-related claim extracted from social media, retrieve the top-10 most relevant abstracts from a corpus of nearly 400,000 scientific abstracts, and (**2**) classify each claim-abstract pair as `SUPPORTS`, `REFUTES`, or `NOT ENOUGH INFORMATION (NEI)`.

Our retrieval pipeline (**Subtask 1**) is a three-stage architecture. First, we combine BM25 (Robertson and Zaragoza, 2009), a fine-tuned dense retriever, and a sparse neural retriever using Reciprocal Rank Fusion (RRF) (Cormack et al., 2009). Second, we train a cross-encoder reranker with a two-phase hard negative mining strategy, leveraging both model uncertainty and relevance judgments. Finally, we apply an adapted RankGPT (Sun et al., 2023), prompting an LLM in a few-shot setting to rerank the top candidates using permutation-based generation informed by cross-encoder scores.

For evidence classification (**Subtask 2**), we eval-

[1] `https://github.com/annamkiepura/ClimateCheck`

Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 293–306
July 31, 2025 ©2025 Association for Computational Linguistics

uate zero- and few-shot LLM prompting, and finetune transformer-based models for multi-class classification of claim-abstract pairs.

Our key contributions are:

- We propose a hybrid multi-stage retrieval framework, incorporating LLM-based permutation generation for reranking, to enhance retrieval effectiveness for automated fact-checking.
- We conduct an evaluation of evidence classification approaches, comparing LLMs under various prompting paradigms against supervised BERT-based classifiers.
- Our system achieves the second-highest performance across both subtasks of the ClimateCheck2025 benchmark, surpassing the official baseline by 53.76% on average across Recall@2, Recall@5, Recall@10, and B-pref.
- We set a new state-of-the-art on the ClimateCheck2025 benchmark in terms of Recall@2.

2 Related Work

Automated fact-checking aims to assess the veracity of claims using evidence, a task traditionally performed by human experts but increasingly addressed with automated methods due to scalability concerns (Nakov et al., 2021). Numerous datasets have been developed to support this research. General-domain resources include FEVER (Wikipedia-based claims) (Thorne et al., 2018), VitaminC (contrastive evidence) (Schuster et al., 2021), LIAR (Wang, 2017), and MultiFC (real-world political/media claims) (Augenstein et al., 2019a). MuMiN further expands this scope to multilingual, multimodal misinformation on social media (Nielsen and McConville, 2022).

Scientific fact-checking, a more specialized subfield, introduces challenges such as complex language, evolving knowledge, and domain-specific reasoning. Key datasets include SciFact (Wadden et al., 2020) (scientific claims and abstracts), HealthVer (Sarrouti et al., 2021) and COVID-Fact (Saakyan et al., 2021) (biomedical), and ClimateViz (climate science) (Su et al., 2025). These corpora underscore the risks of domain-specific misinformation, from harmful medical decisions (Wang et al., 2019) to distorted climate discourse (van der Linden et al., 2017).

The fact-checking process is typically modeled as a pipeline: (1) claim detection (Panchendrarajan and Zubiaga, 2024), (2) check worthiness estimation (Yu et al., 2025), (3) document retrieval (Dey et al., 2025), (4) claim verification via natural language inference (NLI) (Dammu et al., 2024). Some systems also generate explanations, though these face challenges with hallucination (Atanasova et al., 2020). Our work focuses on document retrieval and claim verification.

For retrieval, sparse methods such as BM25 use lexical matching, while dense methods, such as Dense Passage Retrieval (Karpukhin et al., 2020), leverage neural encoders for semantic similarity. Hybrid systems combining both have shown improved performance (Zhang et al., 2024), and retrieval-augmented generation (RAG) models further integrate retrieval with generation for grounded responses (Khaliq et al., 2024).

Claim verification is often framed as an entailment task, with transformer-based models, such as BERT (Devlin et al., 2019), fine-tuned to classify claim-evidence pairs as support, refute, or neutral (Wadden et al., 2022). Prompt-based methods using LLMs offer zero-shot alternatives, though performance varies across models and prompt designs (Chen et al., 2024a).

Despite advances, scientific fact-checking remains challenging due to long-context reasoning, subtle hedging, contradictory evidence, and the need for up-to-date knowledge. LLMs have shown promise as rerankers or classifiers, but often lag behind supervised models in consistency and interpretability (Ghosh et al., 2025). In this work, we investigate how traditional retrieval methods can be effectively integrated with LLMs to leverage their complementary strengths.

3 Dataset

The ClimateCheck dataset (Abu Ahmad et al., 2025a) was prepared by the task's organizers and comprises (i) claims sourced from ClimaConvo (Shiwakoti et al., 2024), DEBAGREEMENT (Pougué-Biyong et al., 2021), Climate-Fever (Diggelmann et al., 2021), MultiFC (Augenstein et al., 2019b), and ClimateFeedback[2], including both real and synthetically generated social media-style content, (ii) a corpus of scientific abstracts from OpenAlex[3] and S2ORC (Lo

[2] https://science.feedback.org/process/

[3] https://openalex.org/

et al., 2020), and (iii) annotated claim-abstract pairs labeled as SUPPORTS, REFUTES, or NOT ENOUGH INFORMATION (NEI). Annotations were produced via TREC-style pooling and reviewed by graduate-level domain experts. Key dataset statistics are summarized in Table 1.

Statistic	Value
Abstract Corpus	
Total # of abstracts	394,269
Mean length (words)	240.93
Min length (words)	1
Max length (words)	6,818
Std dev. of length	232.46
Claims (Train Split)	
Total # of unique claims	252
Mean length (words)	17.76
Min length (words)	3
Max length (words)	43
Std dev. of length	7.50
Claim-Abstract Pairs	
Total # of labeled claim-abstract pairs	1,144
SUPPORT instances	446 (38.99%)
REFUTES instances	241 (21.07%)
NEI instances	457 (39.95%)
Positive instances (SUPPORT + REFUTES)	687 (60.05%)
Relevant Abstracts per Claim	
Mean # of relevant abstracts/claim	2.73
Min # of relevant abstracts/claim	0
Max # of relevant abstracts/claim	5
Std dev. of the # of relevant abstracts/claim	1.68
Claim Relevance Distribution	
# of claims with $\geq$1 supporting abstract	150
# of claims $\geq$1 refuting abstract	101
# of claims with only NEI abstracts	27

Table 1: ClimateCheck dataset statistics.

4 Methodology

Below, we describe our multi-stage pipeline for scientific fact-checking, summarized in Figure 1, and our technical implementation details.

4.1 Subtask 1: Abstract Retrieval

Our approach to **Subtask 1** adopts a retrieve-then-rerank paradigm, inspired by prior multi-stage retrieval systems such as HLART (Zhang et al., 2022), Re2G (Glass et al., 2022), and MST-R (Malviya et al., 2024). In **Stage 1**, we employ a hybrid retrieval setup that combines lexical and neural methods, leveraging their complementary

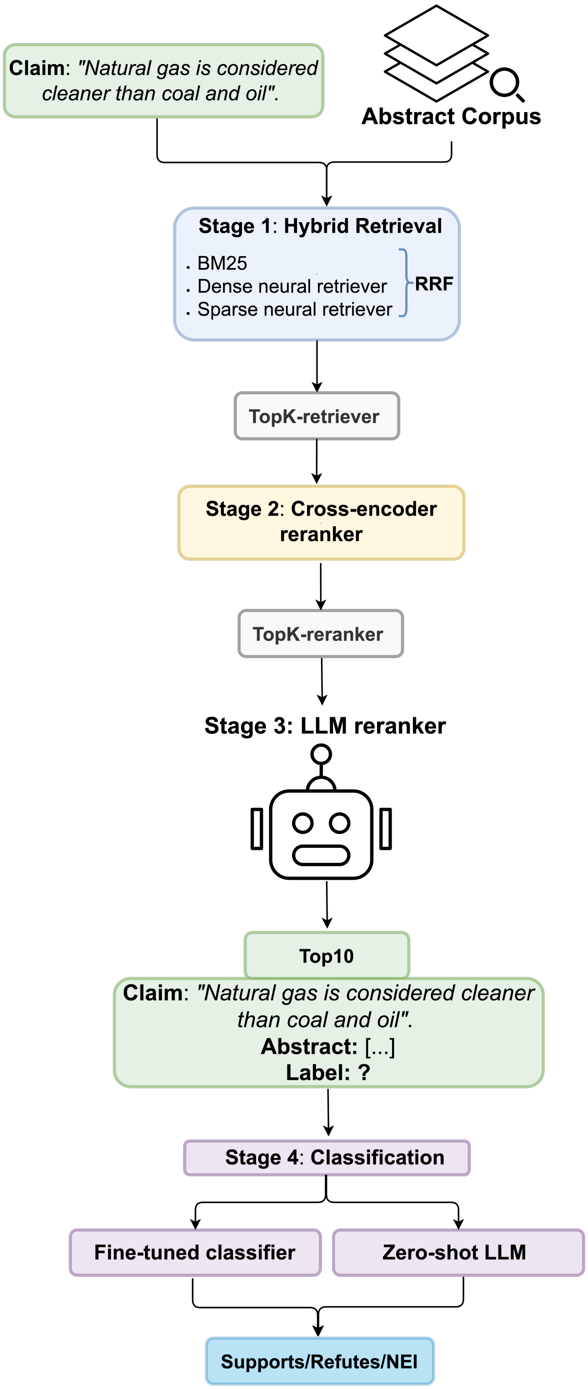

Claim: "Natural gas is considered cleaner than coal and oil".
Abstract Corpus
Stage 1: Hybrid Retrieval
. BM25
. Dense neural retriever
. Sparse neural retriever
RRF
TopK-retriever
Stage 2: Cross-encoder reranker
TopK-reranker
Stage 3: LLM reranker
Top10
Claim: "Natural gas is considered cleaner than coal and oil".
Abstract: [...]
Label: ?
Stage 4: Classification
Fine-tuned classifier
Zero-shot LLM
Supports/Refutes/NEI

Figure 1: Overview of our fact-checking pipeline. A given claim is matched against an abstract corpus using a hybrid retrieval system (**Stage 1**) composed of BM25, dense, and sparse retrievers, fused via Reciprocal Rank Fusion (RRF). Top-ranked abstracts are reranked using a cross-encoder (**Stage 2**), followed by a few-shot LLM-based reranker (**Stage 3**). The final top 10 abstracts are passed to a classification stage (**Stage 4**), where each claim-abstract pair is labeled as SUPPORTS, REFUTES, or NEI using either a fine-tuned classifier or a zero-shot LLM.

strengths: lexical models excel at exact-match precision, while neural models capture semantic similarity. This integration improves overall recall and yields a more diverse set of candidate abstracts, increasing the chances of retrieving relevant evidence that might be overlooked by any single method. Since the combined results originate from heterogeneous retrieval models with non-comparable scoring functions, **Stages 2 and 3** introduce rerankers to normalize and refine the candidate list, enabling coherent and consistent ranking across sources.

4.1.1 Stage 1 - Hybrid Retrieval System

Dense Neural Retriever We fine-tune the BGE-M3 dense retriever model (Chen et al., 2024b) using a triplet loss objective with cosine distance and a margin of 0.3. Each training instance is a triplet consisting of an anchor (the claim), a positive abstract (labeled as `SUPPORTS` or `REFUTES`), and a negative abstract (labeled as `NEI`). The training objective encourages the model to embed the claim closer to the positive abstract than to the negative by a fixed margin in cosine space. Formally, the loss is defined as:

$$\mathcal{L}_{\text{triplet}} = \max\left(0, \cos(\mathbf{q}, \mathbf{d}^-) - \cos(\mathbf{q}, \mathbf{d}^+) + \gamma\right) \tag{1}$$

where $\mathbf{q}$ is the embedding of the claim, $\mathbf{d}^+$ is the embedding of the positive abstract, $\mathbf{d}^-$ is the embedding of the negative abstract, $\cos(\cdot, \cdot)$ denotes cosine similarity, and $\gamma = 0.3$ is the margin.

The model is fine-tuned for 3 epochs with a learning rate of 5×10^{-5}, using a warm-up schedule over the first 10% of training steps. All layers are updated during training, and mixed-precision computation is employed to improve training efficiency. We use a per-device batch size of 2 with gradient accumulation over 32 steps, yielding an effective batch size of 64. Fine-tuning enables the model to better adapt to the scientific domain and capture semantic relationships between claims and evidentiary abstracts more effectively.

After fine-tuning, we precompute dense embeddings for all abstracts in the corpus. At inference time, an input claim is encoded into a dense vector, and similarity scores are computed via the dot product between the claim vector and each abstract embedding. These scores are then used to directly rank the abstracts by relevance.

Sparse Lexical Retriever BM25 is a sparse lexical retriever model based on TF-IDF (Jones, 1972). We build a BM25 index over all corpus abstracts and use it to retrieve most relevant candidates by computing relevance scores between the tokenized claim and each abstract.

Sparse Neural Retriever We utilize a sparse neural retriever based on the pretrained SPLADE-v3 model (Lassance et al., 2024), which encodes queries into high-dimensional sparse vectors by applying a ReLU-activated max pooling over contextualized token logits. Specifically, given contextualized logits $\mathbf{L} \in \mathbb{R}^{T \times V}$ for a query of length T and vocabulary size V, the sparse representation $\mathbf{q} \in \mathbb{R}^V$ is computed as:

$$\mathbf{q}_v = \max_{t=1,\ldots,T} \text{ReLU}(\mathbf{L}_{t,v}) \tag{2}$$

This allows SPLADE to retain the efficiency of inverted index retrieval while incorporating semantic signals from deep transformer architecture. For each input claim, we compute a sparse claim representation and perform retrieval via a sparse dot product against precomputed document vectors of all abstracts.

RRF We combine the ranked outputs of BM25 (Robertson and Zaragoza, 2009), dense, and sparse neural retrievers using Reciprocal Rank Fusion (RRF), a method introduced by Cormack et al. (2009). In RRF, given the rank $r_i(d)$ of document d from retriever i, the final score is computed as:

$$S(d) = \sum_{i=1}^{n} \frac{1}{k_{\text{rrf}} + r_i(d)} \tag{3}$$

where k_{rrf} is a fixed hyperparameter that we set to 60, following the recommendation in Yang et al. (2017). RRF enables effective aggregation of retrieval results from heterogeneous models with non-comparable scoring scales. We apply this hybrid retrieval strategy to select the top-600 candidate abstracts for each claim (see Appendix B for discussion of the top-k choice).

4.1.2 Stage 2 - Cross-Encoder Reranker

At the first reranking stage, we use a cross-encoder model ms-marco-MiniLM-L-6-v2 (Reimers and Gurevych, 2021; Bajaj et al., 2018) trained on the MSMARCO dataset (Wang et al., 2020). Unlike bi-encoders used in **Stage 1**, the cross-encoder jointly encodes the claim and abstract, allowing for richer interaction and more accurate relevance estimation.

Fine-tuning We fine-tune the cross-encoder reranker in two phases using the ClimateCheck annotated dataset, following a curriculum-based learning strategy (Bengio et al., 2009). In the first phase, training examples are constructed by retrieving the top-k candidates (k=200) using the hybrid retrieval system. All truly relevant abstracts (labeled as SUPPORTS or REFUTES in the ground truth) are treated as positive examples. Hard negatives are selected from top-ranked abstracts that are labeled as NEI, while easy negatives are randomly sampled from the remaining NEI abstracts. In the second phase, we use the model trained in the first phase to re-mine more challenging hard negatives. The reranker is then further fine-tuned on this harder set, enabling progressive refinement of its discrimination ability.

We train the model using binary cross-entropy loss, inferred from the scalar output with sigmoid activation and threshold-based label prediction. We use a batch size of 16, a learning rate of 2×10^{-5}, and weight decay of 0.01. Phase 1 includes 3 epochs, followed by 2 additional epochs in phase 2. All experiments are conducted with mixed precision (FP16) (Micikevicius et al., 2018) training enabled for improved efficiency.

Inference At inference time, we retrieve the top-k = 600 candidate abstracts for each test claim using the hybrid retrieval system. The choice of the top-k parameter value used in **Stage 1** is further discussed in Appendix B. These candidates are then reranked using the fine-tuned cross-encoder, and the top-k = 20 are then passed to the **Stage 3** reranker. Different numbers of candidates passed on to the **Stage 3** reranker were not evaluated due to limited resources.

4.1.3 Stage 3 - LLM-based Reranker

The third stage of our pipeline applies an instruction-tuned LLM to rerank the top 20 abstracts produced by the cross-encoder. We use **RankGPT** (Sun et al., 2023) adapted from the official implementation[4], which formulates reranking as a *permutation generation task*. Rather than assigning independent relevance scores (pointwise) or comparing abstract pairs in isolation (pairwise), the model reasons over the entire candidate set holistically and outputs a single ranked list. Given a claim and 20 candidate abstracts, the LLM is prompted in a few-shot setting to generate a permutation $\pi \in \mathbb{S}_N$, where $\pi(i)$ denotes the rank assigned to the i-th abstract. The model is explicitly instructed to order the abstracts from most to least evidentiary, regardless of stance polarity (SUPPORTS or REFUTES). This enables the LLM to model complex interdependencies such as redundancy, diversity, and relative informativeness - capabilities not easily captured by pointwise or pairwise architectures. The resulting LLM-based ranks are converted into normalized scores using:

$$\text{LLM}_{\text{norm}}(d_i) = 1 - \frac{\pi(i) - 1}{N - 1} \tag{4}$$

where N is the number of candidates to rerank and a higher score corresponds to a more evidentiary abstract.

To integrate the LLM's global reasoning with the semantic precision of the cross-encoder, we compute a fused score for each document as:

$$\begin{aligned} \text{score}_{\text{fused}}(d_i) &= \alpha \cdot \text{CE}_{\text{norm}}(d_i) \\ &+ (1 - \alpha) \cdot \text{LLM}_{\text{norm}}(d_i) \end{aligned} \tag{5}$$

where the α parameter balances the contributions of the normalized cross-encoder score $\text{CE}_{\text{norm}}(d_i)$ and the normalized LLM-based score $\text{LLM}_{\text{norm}}(d_i)$. We used $\alpha = 0.4$ as it yielded the best performance. Full ablation results of the value of the α parameter are available in Appendix A. The top 10 abstracts based on the fused scores are then selected as the final ranked evidence set for each claim.

As the LLM, we used GPT-4.1[5] through OpenAI API, with temperature set to 0. Prompting details and the comparison between the zero- and few-shot settings are included in Appendix A.

4.1.4 Retrieval evaluation metrics

We evaluate retrieval performance using several standard metrics: Recall@2, Recall@5, and Recall@10 measure the proportion of relevant abstracts retrieved in the top 2, 5, and 10 positions, respectively. B-Pref (Binary Preference) (Buckley and Voorhees, 2004) quantifies how many relevant items are ranked ahead of non-relevant ones, accounting for incomplete relevance judgments. We also report a composite Retrieval Score, computed as the arithmetic mean of the four preceding metrics.

[4] https://github.com/sunnweiwei/RankGPT

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

4.2 Subtask 2: Stance Classification

To classify the stance of retrieved abstracts (SUPPORTS, REFUTES, or NEI) with respect to the retrieved claims, we explore two approaches: (a) using LLMs in various prompting settings, and (b) training supervised classifiers based on DeBERTa (He et al., 2021) and RoBERTa (Liu et al., 2019), using human-annotated examples in the Climate-Check dataset.

4.2.1 LLM

We experiment with prompting LLMs to classify claim-abstract pairs using both zero-shot and few-shot settings. In the zero-shot setup, the model is directly instructed to assign labels, without any examples provided. In the few-shot variant, we provide examples of annotated pairs to guide the model's reasoning. Additionally, we investigate a two-step classification approach: first, the model predicts whether a given abstract is *evidentiary* (i.e., SUPPORTS or REFUTES) versus *non-evidentiary* (NEI); second, only for the evidentiary stances, a separate model instance predicts the polarity (SUPPORTS vs REFUTES). In the one-step approach, the model is directly prompted to assign one of the three possible labels. In the hybrid approach, a single model instance is instructed to first predict the relevance (evidentiary vs. non-evidentiary), and then the polarity. Full details regarding prompting are available in Appendix C. As the LLM, we used GPT-4.1[6] through OpenAI API, with temperature set to 0.

4.2.2 Supervised fine-tuning

We fine-tune three models, initializing from the following checkpoints: DeBERTa-v3-base-mnli[7], which was trained on the MultiNLI dataset (Williams et al., 2018) consisting of 392,702 NLI hypothesis-premise pairs, DeBERTa-v3-base-scifact[8] and RoBERTa-large-scifact[9], both fine-tuned on the SciFact dataset.

The human-labeled instances were stratified 90/10 into training and validation splits. We freeze the encoder layers, so that only the pooler and classifier layers are updated. To mitigate the mild class imbalance, we employ a custom Trainer that (i) inserts a WeightedRandomSampler so each mini-batch is class-balanced and (ii) replaces the standard cross-entropy with a class-weighted focal loss:

$$\mathcal{L}_{\text{focal}} = -\,\alpha_y \left(1 - p_y\right)^{\gamma} \log p_y \qquad (6)$$

where p_y is the softmax probability of the gold label y, $\alpha_y = 1/f_y$ is the inverse class frequency (normalized so $\sum_c \alpha_c = C$), and $\gamma = 2$ focuses the gradient on hard or minority examples. Training runs for 10 epochs with an effective batch of 32 and a flat learning rate 5×10^{-5}.

4.2.3 Classification evaluation metrics

We report weighted-averaged precision (P), recall (R), and F1-score, which compute metrics for each class and average them according to the number of true instances for the SUPPORTS, REFUTES, and NEI labels. This approach accounts for class imbalance while providing a comprehensive measure of overall system performance.

4.3 Hardware details

All fine-tuning and inference experiments were carried out on the A100 40 GB RAM NVIDIA GPU.

5 Results and Discussion

5.1 Subtask 1: Abstract Retrieval

Alg.	R@2	R@5	R@10	B-Pref	R. Score
B+S+D	0.1447	0.2693	0.3840	0.3102	0.2771
B+S+D+C	0.1882	0.3884	0.5643	0.4270	0.3920
B+S+D+C+L	**0.2309**	**0.4413**	**0.6006**	**0.4818**	**0.4386**
Final baseline	0.1947	0.3047	0.3436	0.2980	0.2853

Table 2: Retrieval results (B=BM25, S=SPLADE, D=Dense, C=Cross-encoder, L=LLM) across retrieval system variants on test dataset. Final baseline refers to the baseline results provided by the task's organizers. Full ablation available in Appendix A.

Retrieval results are shown in Table 2. The initial hybrid retriever, which combines lexical (BM25), sparse neural (SPLADE), and dense (BGE-M3) retrieval methods, achieves a Recall@10 of 0.3840 and a retrieval score of 0.2771. While this baseline benefits from diverse retrieval signals, its ability to rank truly relevant evidence is still limited by the heterogeneous scoring outputs and lack of deeper semantic matching.

Introducing the cross-encoder reranker (B+S+D+C) yields substantial gains across all evaluation metrics. Notably, Recall@10 increases

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

[7] https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli

[8] https://huggingface.co/jedick/DeBERTa-v3-base-mnli-fever-anli-scifact-citint

[9] https://huggingface.co/nikolamilosevic/SCIFACT_xlm_roberta_large

by over 47% (from 0.3840 to 0.5643), while B-Pref improves from 0.3102 to 0.4270. This confirms the effectiveness of cross-encoders in modeling fine-grained semantic relationships between claims and abstracts, particularly in reordering high-recall but noisy candidate sets.

The full pipeline (B+S+D+C+L), which integrates an LLM-based permutation reranker as a final stage, achieves the strongest performance across all metrics. It reaches a Recall@10 of 0.6006 and a B-Pref of 0.4818, corresponding to a final retrieval score of 0.4386. This indicates that the LLM-based reranker provides complementary refinement, likely capturing subtle discourse cues and context-aware relevance signals missed by earlier stages. Improvements are consistent not only in recall-based metrics but also in B-Pref, suggesting that the model is not just retrieving more relevant documents, but also ranking them more coherently with respect to ground truth preferences. Overall, our approach yields an improvement of 53.76% in Retrieval Score over the final baseline published by the shared task's organizers[10].

5.2 Subtask 2: Stance Classification

Version	P	R	F1
LLM prompting			
Few-shots-hybrid	0.6811	0.6835	0.6811
Zero-shot-hybrid	**0.6950**	**0.6973**	**0.6957**
Zero-shot-two-step	0.6780	0.6835	0.6788
Zero-shot-one-step	0.6874	0.6909	0.6842
Supervised fine-tuning			
DeBERTa-v3-base-mnli	0.5468	0.5348	0.5176
DeBERTa-v3-base-scifact	0.5774	0.5285	0.5365
RoBERTa-large-scifact	0.5637	0.5032	0.5098
Final baseline	0.65448	0.62603	0.63148

Table 3: Classification performance across LLM prompting and supervised fine-tuning strategies on the test dataset. Final baseline refers to the baseline results provided by the task's organizers.

Classification results are summarized in Table 3. Among LLM-based strategies, the zero-shot hybrid prompt achieves the highest F1 score of 0.6957, slightly outperforming the few-shot variant (0.6811) and both the one-step and two-step zero-shot setups. This suggests that carefully crafted zero-shot prompts can be as effective - or even more so - than few-shot examples, likely due to reduced prompt length and reduced token-level noise from poorly aligned demonstrations.

The hybrid prompting format, which combines structured instruction with explicit claim-evidence formatting, proves consistently effective across setups. Compared to the two-step approach, where the stance is inferred via intermediate entailment, the one-step and hybrid strategies demonstrate better alignment with the task's categorical stance labels, yielding higher precision and recall. This suggests that direct classification is more robust for LLMs than compositional reasoning pipelines in this context.

Notably, all LLM-based approaches outperform the supervised baselines. The best supervised model (DeBERTa-v3-base fine-tuned on SciFact) achieves an F1 score of 0.5365 - substantially lower than any LLM-based method. This performance gap highlights the limitations of traditional fine-tuning approaches, even when trained on in-domain annotations, and underscores the strength of instruction-tuned LLMs in performing complex stance classification in few- or zero-shot settings.

6 Conclusion

Scientific fact verification poses unique challenges due to complex domain language and the need for precise evidence interpretation. In this work, we introduced a multi-stage retrieval and classification pipeline tailored to these challenges, integrating hybrid retrieval methods, cross-encoder reranking, and LLM-based reasoning modules.

Our experiments on the ClimateCheck benchmark demonstrate consistent improvements across all retrieval metrics, with each additional component - especially LLM-based reranking - contributing meaningfully to performance. In the classification subtask, prompting strategies based on LLMs outperformed traditional fine-tuned models, even when the latter were trained on task-specific human annotations. These findings highlight the flexibility and effectiveness of instruction-tuned LLMs for complex scientific reasoning tasks, especially in data-scarce or rapidly evolving domains.

Overall, our work underscores the importance of combining structured retrieval pipelines with the emergent reasoning abilities of LLMs. Future work could explore more tightly integrated retrieval-generation models, few-shot active learning for stance classification, and methods for im-

[10] The percentage change is calculated as $\left(\frac{0.4386-0.2853}{0.2853}\right) \times 100\% = \left(\frac{0.1533}{0.2853}\right) \times 100\% = 0.5373 \times 100\% = 53.73\%$.

proving the interpretability and trustworthiness of LLM-based decisions in scientific verification contexts.

Limitations

While our system demonstrates strong performance on both retrieval and classification for scientific fact verification, several limitations remain.

First, our retrieval pipeline relies on precomputed document embeddings and staged reranking, which - although effective - can be computationally expensive and may not scale efficiently to real-time or large-scale applications. The use of LLM-based reranking, in particular, introduces latency and resource demands that may be prohibitive in deployment scenarios without high-performance infrastructure.

Second, while prompting-based approaches outperform supervised baselines in our setting, they are sensitive to prompt design and require manual tuning. Our evaluation does not fully explore the robustness of these prompts to variation in phrasing, order, or input format, nor does it address the interpretability of the model's reasoning process.

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.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019a. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019b. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. *Preprint*, arXiv:1909.03242.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

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

Beiduo Chen, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, and Barbara Plank. 2024a. "seeing the big through the small": Can llms approximate human judgment distributions on nli from a few explanations? *Preprint*, arXiv:2406.17600.

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

Gordon V. Cormack, Charles 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.

Preetam Prabhu Srikar Dammu, Himanshu Naidu, Mouly Dewan, YoungMin Kim, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. ClaimVer: Explainable claim-level verification and evidence attribution of text through knowledge graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13613–13627, Miami, Florida, USA. Association for Computational Linguistics.

Jacob Devasier, Rishabh Mediratta, Akshith Putta, and Chengkai Li. 2025. Task-oriented automatic fact-checking with frame-semantics. *Preprint*, arXiv:2501.13288.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Arka Ujjal Dey, Muhammad Junaid Awan, Georgia Channing, Christian Schroeder de Witt, and John Collomosse. 2025. Fact-checking with contextual narratives: Leveraging retrieval-augmented llms for social media analysis. *Preprint*, arXiv:2504.10166.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2021. Climate-fever: A dataset for verification of real-world climate claims. *Preprint*, arXiv:2012.00614.

Bishwamittra Ghosh, Sarah Hasan, Naheed Anjum Arafat, and Arijit Khan. 2025. Logical consistency of large language models in fact-checking. *Preprint*, arXiv:2412.16100.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Rajaram Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2g: Retrieve, rerank, generate. *Preprint*, arXiv:2207.06300.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

Ken Hyland. 1996. Writing without conviction? hedging in science research articles. *Applied Linguistics*, 17(4):433–454.

K. Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. *Preprint*, arXiv:2004.04906.

M. Abdul Khaliq, P. Chang, M. Ma, B. Pflugfelder, and F. Miletić. 2024. Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models. *Preprint*, arXiv:2404.12065.

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

Jiaxin Li and Xiaojun Chang. 2022. Combating misinformation by sharing the truth: a study on the spread of fact-checks on social media. *Information Systems Frontiers*, pages 1–15. Epub ahead of print.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

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.

Yash Malviya, Karan Dhingra, and Maneesh Singh. 2024. Mst-r: Multi-stage tuning for retrieval systems and metric evaluation. *Preprint*, arXiv:2412.10313.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. *Preprint*, arXiv:1710.03740.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. *Preprint*, arXiv:2103.07769.

Dan Saattrup Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. *Preprint*, arXiv:2202.11684.

Haoran Ou, Gelei Deng, Xingshuo Han, Jie Zhang, Xinlei He, Han Qiu, Shangwei Guo, and Tianwei Zhang. 2025. Holmes: Automated fact check with large language models. *Preprint*, arXiv:2505.03135.

Rrubaa Panchendrarajan and Arkaitz Zubiaga. 2024. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Natural Language Processing Journal*, 7:100066.

John Pougué-Biyong, Valentina Semenova, Alexandre Matton, Rachel Han, Aerin Kim, Renaud Lambiotte, and J Doyne Farmer. 2021. Debagreement: A comment-reply dataset for (dis)agreement detection in online debates. In *NeurIPS Datasets and Benchmarks Track (Round 2)*.

Dorian Quelle and Alexandre Bovet. 2024. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, 7.

Nils Reimers and Iryna Gurevych. 2021. The curse of dense low-dimensional information retrieval for large index sizes. *Preprint*, arXiv:2012.14210.

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

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics:*

EMNLP 2021, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ipek Baris Schlicht, Eugenia Fernandez, Berta Chulvi, and Paolo Rosso. 2023. Automatic detection of health misinformation: a systematic review. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–13. Advance online publication.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence. *Preprint*, arXiv:2103.08541.

Shuvam Shiwakoti, Surendrabikram Thapa, Kritesh Rauniyar, Akshyat Shah, Aashish Bhandari, and Usman Naseem. 2024. Analyzing the dynamics of climate change discourse on Twitter: A new annotated corpus and multi-aspect classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 984–994, Torino, Italia. ELRA and ICCL.

Ruiran Su, Jiasheng Si, Zhijiang Guo, and Janet B. Pierrehumbert. 2025. Climateviz: A benchmark for statistical reasoning and fact verification on scientific charts. *Preprint*, arXiv:2506.08700.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. *ArXiv*, abs/2304.09542.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Nicolò Urbani, Sandip Modha, and Gabriella Pasi. 2024. Retrieving semantics for fact-checking: A comparative approach using CQ (claim to question) & AQ (answer to question). In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 37–46, Miami, Florida, USA. Association for Computational Linguistics.

Sander van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. 2017. Inoculating the public against misinformation about climate change. *Global Challenges*, 1(2):1600008.

Ivan Vykopal, Matúš Pikuliak, Simon Ostermann, and Marián Šimko. 2024. Generative large language models in automated fact-checking: A survey. *Preprint*, arXiv:2407.02351.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States. Association for Computational Linguistics.

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.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. Systematic literature review on the spread of health-related misinformation on social media. *Social Science Medicine*, 240:112552.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1253–1256, New York, NY, USA. Association for Computing Machinery.

Yinglong Yu, Hao Shen, Zhengyi Lyu, and Qi He. 2025. Application and optimization of large models based on prompt tuning for fact-check-worthiness estimation. *Preprint*, arXiv:2504.18104.

Haoyu Zhang, Jun Liu, Zhenhua Zhu, Shulin Zeng, Maojia Sheng, Tao Yang, Guohao Dai, and Yu Wang. 2024. Efficient and effective retrieval of dense-sparse hybrid vectors using graph-based approximate nearest neighbor search. *Preprint*, arXiv:2410.20381.

Yanzhao Zhang, Dingkun Long, Guangwei Xu, and Pengjun Xie. 2022. Hlatr: Enhance multi-stage text retrieval with hybrid list aware transformer reranking. *Preprint*, arXiv:2205.10569.

Version	R@2	R@5	R@10	B-Pref	R. Score
Zero-Shot LLM	0.2285	0.4398	0.5961	0.4692	0.4334
Few-Shot LLM	**0.2309**	**0.4413**	**0.6006**	**0.4818**	**0.4387**

Table 4: Comparison of zero-shot and few-shot prompting strategies results for the LLM-based reranker.

> **LLM Prompting Strategy for Passage Ranking (Zero-Shot)**
> **System Prompt:**
> You are given a **CLAIM** and **N PASSAGES**. A passage is *evidentiary* with respect to a claim if it contains information that could either `SUPPORT` or `REFUTE` the claim. Whether it supports or refutes does not matter. Return exactly one line with the passage numbers, most evidentiary first, least evidentiary last. **Output numbers only**.

Figure 2: Prompting strategy for LLM-based passage ranking. Given a claim and a set of passages, the model is instructed to output a permutation of passage indices in decreasing order of evidentiary relevance.

A RankGPT prompting details

To use an LLM as the final reranking stage, we adopt the permutation generation approach (as introduced in RankGPT). It involves instructing an LLM to directly output the permutations of a group of passages. This method ranks passages directly without an intermediate relevance score. To combine the LLM output with the cross-encoder score, we convert the LLM-based ranks into normalized scores, and then compute a fused score incorporating the cross-encoder score for each document, as described in subsection 4.1.3.

The prompt to the LLM is depicted in Figure 2. For each claim, we include the top 20 abstracts retrieved by the cross-encoder. In a few-shots scenario, we additionally incorporate the examples shown in Figure 3. We then produce gold permutations as follows: all evidentiary abstracts (`SUPPORTS` or `REFUTES`) must be ranked higher than any `NEI` abstracts. Except for this rule, the relative order of abstracts is random.

We include the few-shot setting in our final results, as it was demonstrated to yield slightly higher results than the zero-shot setting, as shown in Table 4.

Table 5 contains the results of ablations for the fusion parameter α. As $\alpha = 0.4$ yielded the best overall retrieval performance (as defined by the R. Score), it was included in the final results.

B Full retrieval ablations

Table 6 presents a comprehensive ablation study evaluating different retrieval configurations. Among individual retrievers, SPLADE outperforms BM25 and Dense, particularly in Recall@10 and B-Pref. Adding a cross-encoder (CE) reranker consistently boosts performance across all settings, with SPLADE+CE achieving the best single-retriever reranking results. Combinations of multiple retrievers further improve performance, particularly when fused with the cross-encoder. The best performance is achieved with the full pipeline—BM25 + SPLADE + Dense + CE + LLM—which yields the highest Recall@2, Recall@5, Recall@10, and overall retrieval score.

Table 7 presents additional ablation results focusing on the first-stage retrieval component, comparing different combinations of BM25, Dense, and SPLADE retrievers across varying top-k cutoffs. Individually, SPLADE consistently outperforms BM25 and Dense, especially at lower k, but all three benefit significantly from hybridization. Notably, combining any two retrievers yields substantial gains over individual models. The best overall performance is achieved by the full hybrid—SPLADE + BM25 + Dense—which achieves the highest recall across all k values. These results confirm that hybrid retrieval setups provide more comprehensive and diverse evidence coverage than any single retriever alone. As R@600 is much higher than recall at lower values of k, top 600 abstracts retrieved by the first-stage retrieval component were passed on further to the reranker. Due to time constraints, the effect of setting the value of k for the first-stage retrieval component to 800 and higher on the overall system performance was not evaluated.

C Classification prompting details

For **Subtask 2** (Stance Classification), we tested four different prompting settings.

Few-shots-hybrid involves splitting the classification task into two stages within one prompt. The model is asked to first distinguish between the evidentiary (`SUPPORTS` or `REFUTES`) and non-

Alpha	R@2	R@5	R@10	B-Pref	R. Score
0.0	0.2280	0.3919	0.5728	**0.5016**	0.4236
0.2	**0.2375**	0.4069	0.5884	0.4826	0.4288
0.4	0.2309	**0.4413**	0.6006	0.4818	**0.4386**
0.6	0.1960	0.4151	**0.6044**	0.4521	0.4169
0.8	0.1837	0.3726	0.5795	0.4315	0.3918
1.0	0.1882	0.3884	0.5643	0.4270	0.3920

Table 5: Ablation results for different values of the fusion parameter α, which controls the weighting between LLM-based and cross-encoder (CE) scores in the final reranking step. $\alpha = 0.0$ corresponds to using only the LLM (RankGPT) scores, while $\alpha = 1.0$ corresponds to using only the CE scores.

Algorithm	R@2	R@5	R@10	B-Pref	R. Score
BM25	0.0717	0.1233	0.1803	0.1481	0.1309
Dense	0.0638	0.1123	0.1660	0.1591	0.1253
SPLADE	0.0647	0.1452	0.2190	0.1909	0.1550
BM25 + CE	0.0647	0.1452	0.2190	**0.5044**	0.2333
Dense + CE	0.2001	0.3251	0.4590	0.3715	0.3389
SPLADE + CE	0.1882	0.3511	0.5336	0.4014	0.3686
BM25 + Dense	0.1115	0.2204	0.3080	0.2509	0.2227
SPLADE + Dense	0.1006	0.1796	0.2821	0.2305	0.1982
BM25 + SPLADE	0.1412	0.2522	0.3476	0.2707	0.2529
BM25 + Dense + CE	0.2075	0.3743	0.5493	0.4246	0.3889
SPLADE + Dense + CE	0.1954	0.3683	0.5429	0.4177	0.3811
BM25 + SPLADE + CE	0.1993	0.3639	0.5455	0.4172	0.3815
BM25 + SPLADE + Dense	0.1447	0.2693	0.3840	0.3102	0.2771
BM25 + SPLADE + Dense + CE	0.1882	0.3884	0.5643	0.4270	0.3920
BM25 + SPLADE + Dense + CE + LLM	**0.2309**	**0.4413**	**0.6006**	0.4818	**0.4386**

Table 6: Retrieval performance for various ablation settings. We use RRF to combine the results of multiple models. CE = Cross-Encoder, Dense = dense retriever model.

evidentiary (NEI) abstracts, and then decide if the evidentiary abstracts should be labeled as SUPPORTS or REFUTES. We also provide six examples of claims + three abstracts labeled with respect to their relationship to the corresponding claim. The samples were selected such that each example claim has one supporting, one refuting, and one NEI abstract.

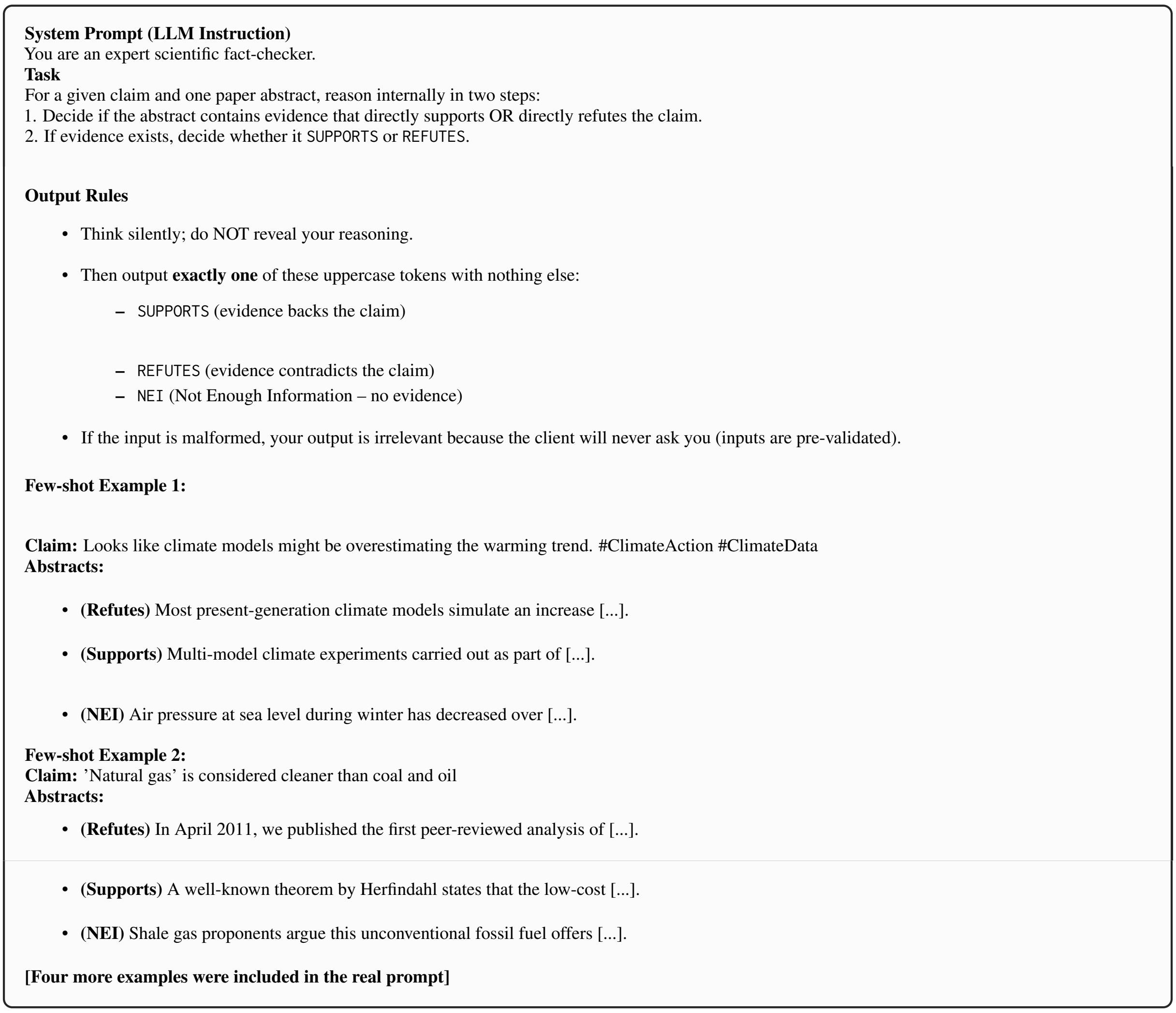
Zero-shot-hybrid involves using the prompt from Figure 2, but without the few-shot examples.

Zero-shot-one-step involves directly asking the model to assign one of the three labels to each claim-abstract pair, as shown in Figure 5.

Zero-shot-two-step involves splitting the classification task into two stages, similarly to **Zero-shot-hybrid**, but using a separate prompt and model instance for each stage, shown in Figure 4.

Algorithm	R@100	R@200	R@400	R@600	R@800
Dense	0.6000	0.6667	0.7333	0.7667	0.7833
BM25	0.3583	0.5083	0.7167	0.8000	0.8667
SPLADE	0.6250	0.7583	0.8083	0.8333	0.8667
BM25 + Dense	0.7000	0.8000	0.8333	0.9083	0.9333
SPLADE + Dense	0.7583	0.7833	0.8833	0.9000	0.9083
SPLADE + BM25	0.7083	0.8167	0.8833	0.9333	0.9500
SPLADE + BM25 + Dense	**0.8417**	**0.8667**	**0.9250**	**0.9583**	**0.9667**

Table 7: Additional ablation results for Stage 1 hybrid retrieval. Dense = dense retriever model.



System Prompt (LLM Instruction)
You are an expert scientific fact-checker.
Task
For a given claim and one paper abstract, reason internally in two steps:
1. Decide if the abstract contains evidence that directly supports OR directly refutes the claim.
2. If evidence exists, decide whether it `SUPPORTS` or `REFUTES`.

Output Rules

- Think silently; do NOT reveal your reasoning.
- Then output **exactly one** of these uppercase tokens with nothing else:
  - `SUPPORTS` (evidence backs the claim)
  - `REFUTES` (evidence contradicts the claim)
  - `NEI` (Not Enough Information – no evidence)
- If the input is malformed, your output is irrelevant because the client will never ask you (inputs are pre-validated).

Few-shot Example 1:

Claim: Looks like climate models might be overestimating the warming trend. #ClimateAction #ClimateData
Abstracts:

- **(Refutes)** Most present-generation climate models simulate an increase [...].
- **(Supports)** Multi-model climate experiments carried out as part of [...].
- **(NEI)** Air pressure at sea level during winter has decreased over [...].

Few-shot Example 2:
Claim: 'Natural gas' is considered cleaner than coal and oil
Abstracts:

- **(Refutes)** In April 2011, we published the first peer-reviewed analysis of [...].
- **(Supports)** A well-known theorem by Herfindahl states that the low-cost [...].
- **(NEI)** Shale gas proponents argue this unconventional fossil fuel offers [...].

[Four more examples were included in the real prompt]

Figure 3: Prompt diagram for the **"few-shots hybrid"** classification configuration. Full prompt included additional four examples, each with one `SUPPORTS`, one `REFUTES`, and one `NEI` abstract.

Two-Step LLM Prompting Strategy for Claim Verification (Zero-Shot)

Step 1: Evidence Detection
System Prompt:

> You are an expert scientific fact-checker.
>
> **Task** Given one claim and one scientific-paper abstract, decide whether the abstract contains evidence that directly supports *or* directly refutes the claim.
>
> **Label Definitions**
> • `EVIDENCE` – The abstract presents data, observations, arguments, or findings that clearly support *or* contradict the claim. Mere topical overlap is insufficient; there must be an evidentiary link.
>
> • `UNKNOWN` – Not enough information. The abstract is off-topic, only tangentially related, or lacks evidence about the claim's truth value.
>
> **Output Rules** 1. Think silently before answering. 2. Output exactly one of the two uppercase tokens, with no extra words, punctuation, or whitespace: `EVIDENCE` or `UNKNOWN` 3. If input is malformed, output `UNKNOWN`. **You must never reveal your reasoning—only the single label.**

Step 2: Polarity Classification
System Prompt:

> You are an expert scientific fact-checker.
>
> **Task** Given one claim and one scientific-paper abstract, decide whether the abstract contains evidence that directly supports OR directly refutes the claim.
>
> **Label Definitions**
> • `SUPPORTS` – The abstract presents data, observations, arguments, or findings that clearly support the claim.
> • `REFUTES` – The abstract presents data, observations, arguments, or findings that clearly contradict the claim.
> (Mere topical overlap is insufficient; there must be an evidentiary link.)
>
> **Output Rules** 1. Think silently before answering. 2. Then output exactly one of the two lowercase tokens, with no extra words,
>
> punctuation, or whitespace: `SUPPORTS` or `REFUTES` 3. If the inputs are missing or malformed, output `UNKNOWN`. **You must never reveal your reasoning—only the single label.**

Note: Abstracts labeled as `UNKNOWN` in Step 1 are not passed to Step 2.

Figure 4: Two-step prompting strategy used for **"zero-shot-two-step"** classification configuration. Step 1 filters out non-evidentiary abstracts, and Step 2 assigns polarity labels (`SUPPORTS` or `REFUTES`) to the evidentiary ones.

One-Step LLM Prompting Strategy for Claim Verification (Zero-Shot)

System Prompt:

> You are an expert scientific fact-checker.
> Given a claim and a paper abstract, reply with exactly one of: `supports | refutes | not enough information`

Figure 5: Prompt for the **"zero-shot-one-step"** classification configuration.