+VeriRel: Verification Feedback to Enhance Document Retrieval for Scientific Fact Checking

Xingyu Deng
xdeng37@sheffield.ac.uk
University of Sheffield
Sheffield, UK

Xi Wang
xi.wang@sheffield.ac.uk
University of Sheffield
Sheffield, UK

Mark Stevenson
mark.stevenson@sheffield.ac.uk
University of Sheffield
Sheffield, UK

Abstract

Identification of appropriate supporting evidence is critical to the success of scientific fact checking. However, existing approaches rely on off-the-shelf Information Retrieval algorithms that rank documents based on relevance rather than the evidence they provide to support or refute the claim being checked. This paper proposes +VeriRel which includes verification success in the document ranking. Experimental results on three scientific fact checking datasets (SciFact, SciFact-Open and Check-Covid) demonstrate consistently leading performance by +VeriRel for document evidence retrieval and a positive impact on downstream verification. This study highlights the potential of integrating verification feedback to document relevance assessment for effective scientific fact checking systems. It shows promising future work to evaluate fine-grained relevance when examining complex documents for advanced scientific fact checking.

CCS Concepts

• **Information systems** → **Specialized information retrieval**.

Keywords

Evidence retrieval, Scientific fact checking, Document ranking

ACM Reference Format:
Xingyu Deng, Xi Wang, and Mark Stevenson. 2025. +VeriRel: Verification Feedback to Enhance Document Retrieval for Scientific Fact Checking. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3746252.3760822

1 Introduction

Automated scientific fact checking is the process of verifying or refuting scientific claims using peer-reviewed research as evidence. The problem has become increasingly important given the rapid growth rate of scientific knowledge and the associated challenge of keeping up to date. Increasing interest in the problem is demonstrated by the development of novel approaches and release of datasets, e.g., [8, 13, 17, 24, 26, 28].

A critical stage of the fact checking process is the identification of evidence against which a claim can be evaluated [3]. This problem normally involves searching for documents containing this evidence within a large corpus, such as a repository of scientific publications. A distinction can be made between "evidential" and semantically relevant documents. The first contains evidence regarding the claim's correctness, while the second includes information related to the claims, but may or may not contain useful evidence. For example, given the claim "*Ibuprofen is frequently used to treat headaches in COVID-19 patients.*", the statement "*Ibuprofen can be used for headache treatment*" is relevant but not evidential. However, "*Paracetamol is the most commonly used medicine for COVID-19 for any symptom*" is both relevant and evidential. Previous work has demonstrated that retrieving relevant but non-evidential documents can be harmful to the fact checking process. For example, Sauchuk et al. [18] reported that combining evidential documents identified by a perfect retriever with a single relevant but non-evidential document produced a 17.2% drop in fact checking performance.

Previous approaches to optimising evidential information have focused on sentence-level evidence selection. Zheng et al. [32] employed joint learning but their tight coupling of retrieval and verification risks overfitting and limited generalisability. Others explored sequential retriever refinement [6, 30] which requires gold sentence-level evidence for supervision, limiting applicability beyond annotated data. A significant limitation of previous approaches is that they assume the documents containing the evidential information are already provided, ignoring the fact that this information is not normally available. Despite its importance, integrating verification feedback into document-level retrieval remains an unexplored problem.

This study addresses this gap to advance document retrieval for scientific fact checking using joint estimation of semantic relevance and downstream verification success of documents. The proposed approach is trainable on arbitrary claim–document pairs using automatically derived verification feedback, thereby enabling broad applicability and scalable training. The main contributions of this paper are:

• Propose *+VeriRel*, a trainable reranker that effectively integrates verification feedback for document ranking. Results show consistent leading performance compared to state-of-the-art baselines across three datasets.
• Explore the factors that can affect the scalability and domain generalisability of *+VeriRel*, which suggests a novel way to train robust document reranking models for scientific fact checking.

2 Methodology

2.1 Task definition

Scientific fact checking systems aim to assess a claim, c, based on the information contained within a corpus of documents, D. Given

CIKM '25, November 10–14, 2025, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3760822

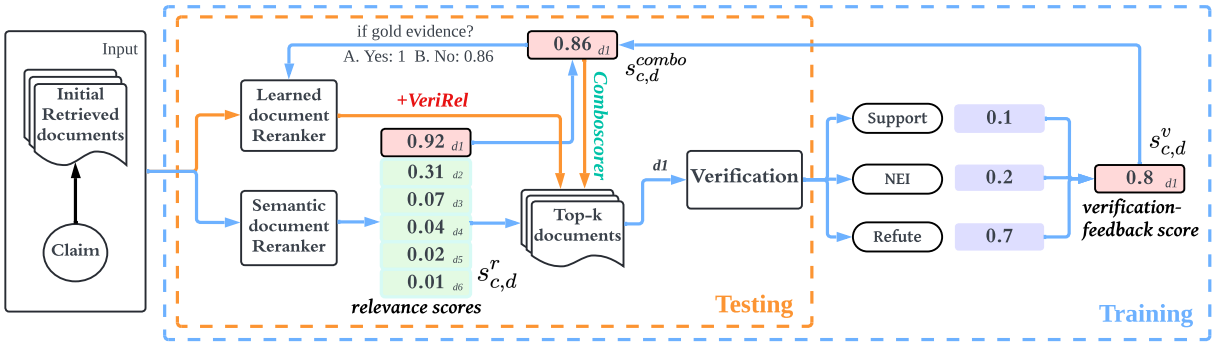

Figure 1: Pipeline to leverage verification feedback to enhance document retrieval. The blue pathway presents the flow of producing joint scores for *+VeriRel* model training. The orange pathway shows the test of ranked documents (1) by calculated scores and (2) by trained *+VeriRel*.

this information, a fact checking system, FC, identifies a set of documents, $D' \subset D$, and predicts a label for each by a classification model, with labels drawn from a pre-defined set, $\mathcal{L}$, i.e.

$$FC(c, D) \rightarrow \{(d, l) : d \in D', l \in \mathcal{L}\} \tag{1}$$

Each label is an assessment of how the information within the document relates to the veracity of the claim. A commonly used a set labels is $\mathcal{L} = \{\texttt{SUPPORT}, \texttt{REFUTE}, \texttt{NOT ENOUGH INFORMATION}\}$.

This process is normally carried out in a two-stage pipeline: evidence retrieval and verification. The first of these, $Retrieval(\cdot)$, treats c as a query and retrieves D', the top k ranked documents from D:

$$D' = Retrieval(c, D, k) \tag{2}$$

The verification stage is then applied to each document $d \in D'$ to determine its label:

$$V(c, d) = \underset{l \in \mathcal{L}}{argmax}(P(l|c, d)) \tag{3}$$

where $P(l|c, d)$ is an estimate of the probability of label l given the claim and document. The verifier chooses the label with the highest probability score.

2.2 +VeriRel

Document retrieval typically involves two steps: initial retrieval for high recall, followed by reranking to identify top-relevant documents [5]. This pipeline allows scientific fact checking systems to obtain semantically relevant documents effectively. However, as noted in Section 1, semantic relevance alone may introduce non-evidential noise. Since the verification stage estimates how likely a document supports, refutes, or lacks information for a claim, it is therefore hypothesised that *incorporating verification feedback into retrieval can enhance the identification of truly evidential documents.*

+VeriRel combines semantic relevance and verification feedback to improve document ranking. A document's verification usefulness as feedback $s^{v}_{c,d}$ is defined as the sum of support and refute probabilities produced by the verifier (Eq. 3), while the predicted probability of the NOT ENOUGH INFORMATION label is intentionally ignored:

$$s^{v}_{c,d} = P(\texttt{SUPPORT}|c, d) + P(\texttt{REFUTE}|c, d) \tag{4}$$

Semantic relevance, $s^{r}_{c,d}$, is computed using an existing ranking model. The final ranking score is a weighted combination that approximates a non-learned ranking strategy to take account of both the semantic and evidential value of a document:

$$s^{combo}_{c,d} = \alpha \times s^{v}_{c,d} + (1 - \alpha) \times s^{r}_{c,d}, \ \alpha \in [0, 1] \tag{5}$$

However, this formulation requires explicit verifier outputs for every document which limits its practicality for large-scale deployment. To address this, *+VeriRel* approximates the joint signal:

$$s^{+VeriRel} = f(c, d) \approx s^{combo}_{c,d} \tag{6}$$

where $f(c, d)$ is a function that models the relationship between a claim c and a document, using a trainable end-to-end model to learn from calculated $s^{combo}_{c,d}$ and gold evidence.

As shown in Figure 1, the blue pathway generates $s^{combo}_{c,d}$ as training signals from semantic relevance and verification feedback, which are then used to train *+VeriRel*. The orange pathway presents the evaluation of ranked documents (1) by the learned model *+VeriRel*, and (2) by $s^{combo}_{c,d}$ directly (denoted as *ComboScorer*).

3 Experiments

This section discusses a series of experiments to (1) address our preliminary study that validates our assumption about the usefulness of verification feedback and (2) train and evaluate our proposed *+VeriRel* reranking model in a scientific fact checking system. Our code and data are publicly available.[1]

3.1 Datasets

Three publicly available datasets were used to provide evidence-annotated corpora:

SciFact [24] contains 809/300 claims for training/validation and 5,183 scientific abstracts from S2ORC [12]. SciFact also includes a test set consisting of 300 claims. Gold evidence for these has not been publicly released but can be inferred from the SciFact-Open dataset [26] to produce a dataset consisting of 279 claims which was used to evaluate the approaches.

[1] https://github.com/xingyu-deng/VeriRel

SciFact-Open [26] is an extension of SciFact which reuses 279 of its test claims while expanding the corpus to 500K abstracts with new evidence annotations. In this study, a distilled dataset comprising evidence only from newly incorporated documents is used to assess the generalisability.

Check-COVID [28] comprises 1,504 COVID-19 related claims and 347 scientific documents. The full dataset is used for evaluation.

All models were trained only using the SciFact training and validation sets, then evaluated using the SciFact, SciFact-Open and Check-COVID evaluation data.

3.2 Model Selection

Reranker: The BM25[16] and monoT5-3B[14] pipeline was adopted as a baseline since it was the best-performing retrieval system for SciFact [11, 19], as reported through the BEIR leaderboard [21], and effective for scientific fact checking [15, 23, 26, 27, 29]. Both monoT5-3B and its MS MARCO MED variant, denoted monoT5-3B(Med), were applied given the biomedical nature of the datasets.

Claim Verifier: MultiVerS [27] was used since it is the current state-of-the-art claim verifier on SciFact and SciFact-Open according to the task leaderboard and published results [26, 27]. MultiVerS uses the Longformer model [2] which enables the processing of long documents to cover abstract-level text and avoid information loss. MultiVerS is initialised with the checkpoint [27] trained on three datasets: FEVER [22], PubMedQA [7] and Evidence Inference [4, 9].

3.3 Negative sampling

All documents in the datasets other than gold evidence are unannotated and treated as being labelled `NOT ENOUGH INFORMATION`. The number of negative samples plays a key role in balancing generalisation and overfitting [26]. In practice, verification models are commonly trained with a substantial number of negative samples to improve robustness and achieve higher overall verification performance, particularly in terms of F1 score [10, 24, 27, 31]. For instance, MultiVerS uses 20 negative samples for each positive one during training [27].

To systematically investigate how verification supervision affects document reranking, the verifier was trained with different numbers of negative documents (N = 5, 10, 20), randomly sampled from the top-100 documents ranked by BM25 in the SciFact train set. This verifier, used to provide feedback ($s^v_{c,d}$), is referred to as V-MultiVerS, to distinguish it from the off-the-shelf MultiVerS used in later verification evaluations. These signals contribute to supervision $s^{combo}_{c,d}$ for *+VeriRel*, allowing examination of how negative sampling strategies affect the quality and generalisability of retrieval training.

Table 1 shows the effect of different negative sampling strategies (N) on verification performance. It can be observed that using more negative documents (N=20) improves precision (i.e., specificity) and achieves higher verification performance (as F1 score), while fewer (N=5) yield better recall and generalisation.

It is possible that using fewer negative samples (i.e. lower values for N) exposes the verifier to more informative borderline cases, leading to richer supervision signals for document reranking. This may help reduce overfitting to trivial non-evidence examples and improve cross-domain generalisation when training *+VeriRel*.

Table 1: Verification performance on SciFact.

Model	#N	Precision	Recall	F1
V-MultiVerS	20	**62.16**	72.52	**66.94**
	10	57.71	72.52	64.27
	5	41.45	**77.48**	54.00

3.4 +VeriRel Configurations

Training s^{combo} consists of (1) a semantic relevance score ($s^r_{c,d}$), which is calculated using a sigmoid function over the output of monoT5-3B, and (2) a verification feedback score ($s^v_{c,d}$) from V-MultiVerS. Different values for the hyperparameter α in Eq. 5 were experimented with. The best-performing was found to vary across negative sampling strategies but values around 0.5 consistently performed well. To ensure fair comparison the simple and intuitive setting of $\alpha = 0.5$ was adopted for all #N settings.

To compute $s^{+VeriRel}$ in Eq. 6, a pre-trained model SciBERT [1], effective in scientific domains [8, 20, 24], was used to calculate a joint relevance score in the range of [0,1]. During model training, the supervision label of document d is set to 1 if d is gold evidence or $s^{combo}_{c,d}$ otherwise. Model training data consisted of the top 20 documents ranked by $s^{combo}_{c,d}$ in the SciFact train set combined with any gold evidence not otherwise included.

3.5 Evaluation Metrics

Recall@k assesses the proportion of relevant evidence included in the top k results:

$$Recall@k = \frac{N(retrieved@k)}{N(gold\_evidence)} \tag{7}$$

where $N(retrieved@k)$ and $N(gold\_evidence)$ are, respectively, the number of gold evidence documents in the top k of ranked list and in the entire corpus.

Verification performance is measured using the standard metrics provided from SCIVER shared task (precision, recall and F1) [25]. The recreated SciFact is denoted as SciFact(offline) to distinguish it from SciFact(leaderboard). For SciFact(leaderboard), we use the unprocessed dataset and evaluate by submitting to the provided leaderboard.[2]

SciFact-Open and Check-COVID are excluded from verification evaluation due to MultiVerS's limited generalisability, as it performs poorly even when providing gold documents in the oracle setting.

4 Results

Evaluation compares retrieval effectiveness of documents ranked (1) by $s^{combo}_{c,d}$ directly (denoted as ComboScorer) and (2) by a trained ranking model *+VeriRel*, using Recall@k across SciFact, SciFact-Open, and Check-COVID.

Table 2 shows that both approaches outperform the strong baselines, particularly at lower values of k. In addition, an ablation study excluded the verification signal by setting $\alpha = 0$, denoted by "baseline ($\alpha = 0$)". These results highlight the benefit of incorporating verification feedback into document scoring, either directly at inference or as supervision for reranker training.

On SciFact, both ComboScorer and *+VeriRel* outperform the state-of-the-art monoT5-3B across all Recall@k values. Among

[2]https://leaderboard.allenai.org/scifact

Table 2: Performance of baselines, ComboScorer and *+VeriRel* in Recall@k with cut-off k ranges from 50 to 1.

Method		Recall - SciFact						Recall - SciFact-Open						Recall - Check-COVID					
		R@50	R@20	R@10	R@5	R@3	R@1	R@50	R@20	R@10	R@5	R@3	R@1	R@50	R@20	R@10	R@5	R@3	R@1
BM25		73.68	67.94	61.24	55.50	48.33	35.41	59.76	43.82	31.87	23.51	16.33	7.57	87.91	81.96	75.02	67.59	61.35	46.18
monoT5-3B		90.91	87.56	85.65	78.47	70.33	55.02	87.25	72.11	58.96	39.44	29.88	11.16	95.84	93.16	89.49	82.06	74.93	58.28
monoT5-3B(Med)		91.39	87.56	85.17	78.95	70.81	55.50	85.66	71.31	57.77	41.04	28.69	10.36	95.54	93.26	89.20	81.86	74.93	57.88
Ours	#N																		
ComboScorer	20	**92.82**	**89.47**	85.65	80.38	**77.51**	**61.72**	87.65	72.11	62.15	43.82	30.68	11.55	95.74	93.76	90.68	83.75	76.51	56.39
	10	92.34	89.00	**87.08**	**82.30**	76.56	60.29	88.84	74.10	**63.75**	44.22	33.07	11.16	**96.13**	94.05	**91.58**	**84.64**	76.71	59.56
	5	92.34	88.52	85.65	81.34	73.21	55.98	**91.63**	**76.49**	59.36	**47.01**	**35.86**	**11.95**	**96.13**	**94.55**	91.48	84.04	**77.80**	**61.84**
baseline ($\alpha = 0$)	–	90.91	88.52	85.65	78.95	73.21	57.89	84.46	70.92	55.38	35.46	24.70	9.96	96.73	91.50	83.66	71.24	60.13	43.14
+VeriRel	20	**91.87**	89.47	**87.08**	79.90	75.12	60.77	85.26	70.52	56.97	40.24	27.89	9.96	96.73	92.81	90.85	78.43	69.28	48.37
	10	**91.87**	89.47	85.65	80.38	75.12	61.72	86.45	71.31	58.17	40.64	27.09	10.76	96.73	94.12	91.50	81.05	71.24	53.59
	5	**91.87**	**90.43**	**87.08**	**82.30**	**75.60**	**62.20**	**87.65**	**72.91**	58.57	**41.43**	28.69	**11.55**	**97.39**	**96.08**	**92.81**	**86.93**	**80.39**	55.56

Table 3: Verification performance by inputting document retrieved by baseline and by proposed *+VeriRel* with $N = 5$.

Model	Verification performance			Retrieval performance
+ MultiVerS	P	R	F1	Recall@k
Top 10	SciFact(offline)			Recall@10
monoT5-3B	74.52	55.98	63.93	86.60
+VeriRel	**75.88**	**61.72**	**68.07**	**88.04**
Top 5	SciFact(offline)			Recall@5
monoT5-3B	**73.03**	62.20	67.18	80.38
+VeriRel	72.48	**65.55**	**68.84**	**84.21**
Top 3	SciFact(offline)			Recall@3
monoT5-3B	76.43	57.42	65.57	73.68
+VeriRel	**77.78**	**63.64**	**70.00**	**78.95**
Top 10	SciFact(leaderboard)			
monoT5-3B	**73.83**	**71.17**	**72.48**	Not accessible
+VeriRel	**73.83**	**71.17**	**72.48**	Not accessible
Top 5	SciFact(leaderboard)			
monoT5-3B	74.88	69.82	72.26	Not accessible
+VeriRel	**75.12**	**70.72**	**72.85**	Not accessible
Top 3	SciFact(leaderboard)			
monoT5-3B	**77.04**	68.08	72.25	Not accessible
+VeriRel	75.86	**69.37**	**72.47**	Not accessible

ComboScorer variants, including more negative samples (N=20) yields the highest early retrieval scores (e.g., Recall@1 = 61.72, Recall@3 = 77.51), suggesting that more negatives help improve document-level precision. Interestingly, when ComboScorer is used to supervise *+VeriRel* training, the optimal negative sampling setting shifts: *+VeriRel* (N=5) achieves the best retrieval performance, surpassing all configurations including its own supervision source, with Recall@1 = 62.20, Recall@3 = 75.60 and Recall@5 = 82.30. This contrast highlights a key insight: although high-N verification signals yield stronger precision in isolation, lower-N settings provide more informative supervision for training a general-purpose reranker.

To evaluate generalisability, all approaches were also evaluated using SciFact-Open and Check-COVID. The *+VeriRel* variant trained with N=5 demonstrates the most consistent performance across both datasets, outperforming other configurations at all Recall@k levels. Notably, Check-COVID presents a greater challenge due to its distinct domain and evolving terminology, yet *+VeriRel*(N=5) achieves the highest scores with substantial gains over all baselines, indicating strong robustness to distribution shift. These results suggest that supervision from fewer hard negative documents may enhance generalisability by exposing the model to more diverse and ambiguous training signals. While high-N verifiers offer stronger in-domain signals, they may lead to overfitting to domain-specific patterns, reducing robustness under distribution shift. One limitation of the Check-COVID dataset is that each claim is annotated with evidence from a single source document, while relevant content in other documents remains unlabelled. This incomplete annotation penalises correct retrievals, particularly at Top-1, where relevant but unlabelled documents are treated as false positives. As a result, while *+VeriRel* improves overall retrieval performance, its performance on Check-COVID is still underestimated.

For evaluation of downstream verification, documents reranked by *+VeriRel*(N=5) improve MultiVerS performance over monoT5-3B across top-10, top-5, and top-3 settings, as shown in Table 3. For instance, on SciFact (offline), F1 improves from 65.57% to 70.00% at top-3 input. On the SCIVER leaderboard[2], *+VeriRel* matches or slightly exceeds monoT5-3B in verification F1 while requiring fewer documents. These gains support that improved reranking quality translates directly to better verification. In summary, *+VeriRel* improves both document retrieval and downstream verification, demonstrating the benefit of integrating verification feedback into document relevance estimation.

5 Conclusion

This study presents *+VeriRel*, a document reranking approach that enhances scientific fact checking by leveraging feedback from verification stages. By integrating a verification reward model into the ranking process, *+VeriRel* prioritises more relevant and evidential documents and consistently improves retrieval accuracy. Experiments demonstrate its strong generalisability, particularly on large, diverse, and previously unseen corpora, exemplified by fast-evolving domains such as COVID-19. The use of downstream verification feedback as an automated relevance feedback mechanism enables more robust and effective document retrieval, which is particularly crucial in scientific domains that demand precision and high-quality evidence. Results presented here demonstrate the benefit of decoupling the training of the verification reward model and the final claim verifier. The reward model should be trained with fewer negative samples to support generalisable evidence identification. Once the reranker is optimised using this feedback, a separate, task-specific verifier can be trained for inference. The improvements from integrating verification supervision may transfer to general fact checking. Future work could adapt our training paradigm to open-domain datasets, especially where dependence on search APIs limits retrieval quality and transparency.

GenAI Usage Disclosure

This manuscript has benefited from the use of Generative AI (ChatGPT) to improve the quality of text produced by the authors. The authors remain responsible for the content.

References

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. 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)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3615–3620. https://doi.org/10.18653/v1/D19-1371

[2] Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020).

[3] Xingyu Deng, Xi Wang, and Mark Stevenson. 2025. The Next Phase of Scientific Fact-Checking: Advanced Evidence Retrieval from Complex Structured Academic Papers. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)* (Padua, Italy) *(ICTIR '25)*. Association for Computing Machinery, New York, NY, USA, 436–448. https://doi.org/10.1145/3731120.3744614

[4] Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020. Evidence Inference 2.0: More Data, Better Models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii (Eds.). Association for Computational Linguistics, Online, 123–132. https://doi.org/10.18653/v1/2020.bionlp-1.13

[5] Kailash A Hambarde and Hugo Proenca. 2023. Information retrieval: recent advances and beyond. *IEEE Access* (2023).

[6] Xuming Hu, Zhaochen Hong, Zhijiang Guo, Lijie Wen, and Philip Yu. 2023. Read it twice: Towards faithfully interpretable fact verification by revisiting evidence. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2319–2323.

[7] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. 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)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 2567–2577. https://doi.org/10.18653/v1/D19-1259

[8] Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking for Public Health Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 7740–7754. https://doi.org/10.18653/v1/2020.emnlp-main.623

[9] Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring Which Medical Treatments Work from Reports of Clinical Trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 3705–3717. https://doi.org/10.18653/v1/N19-1371

[10] Xiangci Li, Gully A Burns, and Nanyun Peng. 2021. A Paragraph-level Multi-task Learning Model for Scientific Fact-Verification.. In *SDU@ AAAI*.

[11] Qi Liu, Bo Wang, Nan Wang, and Jiaxin Mao. 2025. Leveraging passage embeddings for efficient listwise reranking with large language models. In *Proceedings of the ACM on Web Conference 2025*. 4274–4283.

[12] 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*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 4969–4983. https://doi.org/10.18653/v1/2020.acl-main.447

[13] Isabelle Mohr, Amelie Wührl, and Roman Klinger. 2022. CoVERT: A Corpus of Fact-checked Biomedical COVID-19 Tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nichakrabarty2018robustcoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 244–257. https://aclanthology.org/2022.lrec-1.26

[14] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 708–718. https://doi.org/10.18653/v1/2020.findings-emnlp.63

[15] Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. Scientific Claim Verification with VerT5erini. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, Eben Holderness, Antonio Jimeno Yepes, Alberto Lavelli, Anne-Lyse Minard, James Pustejovsky, and Fabio Rinaldi (Eds.). Association for Computational Linguistics, online, 94–103. https://aclanthology.org/2021.louhi-1.11

[16] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp* 109 (1995), 109.

[17] 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*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 3499–3512. https://doi.org/10.18653/v1/2021.findings-emnlp.297

[18] Artsiom Sauchuk, James Thorne, Alon Halevy, Nicola Tonellotto, and Fabrizio Silvestri. 2022. On the role of relevance in natural language processing tasks. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1785–1789.

[19] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 14918–14937. https://doi.org/10.18653/v1/2023.emnlp-main.923

[20] Neset Tan, Trung Nguyen, Josh Bensemann, Alex Peng, Qiming Bao, Yang Chen, Mark Gahegan, and Michael Witbrock. 2023. Multi2Claim: Generating Scientific Claims from Multi-Choice Questions for Scientific Fact-Checking. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 2652–2664. https://doi.org/10.18653/v1/2023.eacl-main.194

[21] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663* (2021).

[22] 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)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 809–819. https://doi.org/10.18653/v1/N18-1074

[23] Juraj Vladika and Florian Matthes. 2023. Scientific Fact-Checking: A Survey of Resources and Approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 6215–6230. https://doi.org/10.18653/v1/2023.findings-acl.387

[24] 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)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 7534–7550. https://doi.org/10.18653/v1/2020.emnlp-main.609

[25] David Wadden and Kyle Lo. 2021. Overview and Insights from the SCIVER shared task on Scientific Claim Verification. In *Proceedings of the Second Workshop on Scholarly Document Processing*, Iz Beltagy, Arman Cohan, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Keith Hall, Drahomira Herrmannova, Petr Knoth, Kyle Lo, Philipp Mayr, Robert M. Patton, Michal Shmueli-Scheuer, Anita de Waard, Kuansan Wang, and Lucy Lu Wang (Eds.). Association for Computational Linguistics, Online, 124–129. https://aclanthology.org/2021.sdp-1.16

[26] David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-Open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 4719–4734. https://doi.org/10.18653/v1/2022.findings-emnlp.347

[27] 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*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 61–76. https://doi.org/10.18653/v1/2022.findings-naacl.6

[28] Gengyu Wang, Kate Harwood, Lawrence Chillrud, Amith Ananthram, Melanie Subbiah, and Kathleen McKeown. 2023. Check-COVID: Fact-Checking COVID-19 News Claims with Scientific Evidence. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 14114–14127. https://doi.org/10.18653/v1/2023.findings-acl.888

[29] Amelie Wührl and Roman Klinger. 2021. Claim Detection in Biomedical Twitter Posts. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii (Eds.). Association for Computational Linguistics, Online, 131–142. https://doi.org/10.18653/v1/2021.bionlp-1.15

[30] Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2023. From Relevance to Utility: Evidence Retrieval with Feedback for Fact Verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6373–6384. https://doi.org/10.18653/v1/2023.findings-emnlp.422

[31] Zhiwei Zhang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. 2021. Abstract, Rationale, Stance: A Joint Model for Scientific Claim Verification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 3580–3586. https://doi.org/10.18653/v1/2021.emnlp-main.290

[32] 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*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 9274–9281. https://aclanthology.org/2024.findings-acl.551