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

Anonymous ACL submission

Abstract

A common fact checking pipeline relies on retrieving evidential documents to guarantee effective verification. However, existing studies ignore the value of improving document retrieval effectiveness. This study improves off-the-shelf semantic-based document retrievers for scientific fact checking with the consideration of the verification success factor to justify the actual relevance. We first develop an ideal document relevance scorer, ComboScorer, to validate the effectiveness of the validation feedback combined ranking model. Then, we propose +VeriRel to approximate the joint feedback for automatic relevance assessment. The experimental results on three scientific fact checking datasets, SciFact, SciFact-Open and Check-Covid, show consistently leading performance of +VeriRel on retrieval effectiveness and its positive impact on the downstream verification. The findings of this study also highlight the potential of integrating verification feedback to document relevance assessment for effective scientific fact checking systems.

1 Introduction

Interest in scientific fact checking, the task of verifying scientific claims using peer-reviewed research as evidence, has recently increased as demonstrated by developing novel approaches and release of datasets, e.g., (Kotonya and Toni, 2020; Sarrouti et al., 2021; Mohr et al., 2022; Wang et al., 2023; Lu et al., 2023). Document retrieval is an essential component in scientific fact checking. The information required to verify or refute scientific claims is often contained within only a few documents (Vladika and Matthes, 2023). This is in contrast to open-domain fact checking where the necessary information may be available from multiple sources (Thorne et al., 2018). Indeed, open-domain fact checking tasks often rely on search engines (Zeng et al., 2021; Guo et al., 2022; Schlichtkrull et al., 2023) such as MediaWiki (Hanselowski et al., 2018) and Google search APIs (Chakrabarty et al., 2018).

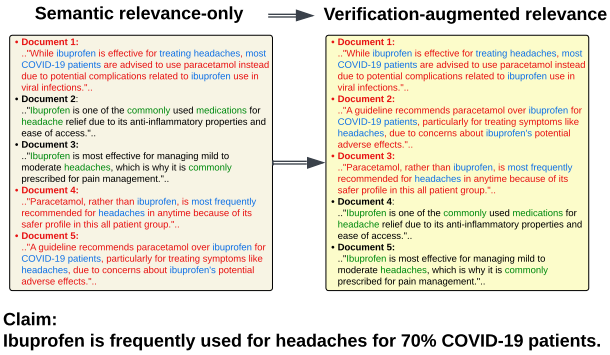

Figure 1: Ranking optimisation example. The red colour indicates the evidential text.

Existing approaches to scientific fact checking rank potential evidence based on its semantic relevance to a claim (Vladika and Matthes, 2023; Wadden et al., 2020; Pradeep et al., 2021; Li et al., 2021; Zhang et al., 2021; Wadden et al., 2022b; Wührl and Klinger, 2021) rather than whether it contains evidence or not. For example, for the claim "*Ibuprofen is frequently used for headaches for 70% COVID-19 patients.*", the statement "*Ibuprofen can be used for headache treatment*" is relevant but can not effectively justify the correctness of the claim. In this circumstance, a fact checking system needs a statement like "*Paracetamol is the most commonly used medicine for COVID-19 for any symptom*" for verification. Figure 1 shows two retrieved document lists with respect to the above claim, one relies on relevance only (left) and another takes the verification success factor into account (right). It shows that top-ranked documents that are semantically relevant do not necessarily contribute to claim verification.

A mixture of evidential and non-evidential documents can produce inaccurate fact checking conclusions. For example, Sauchuk et al. (2022) reported that using a retrieval method that recalls all required evidence for a claim but mixing a single

piece of negative evidence with the gold evidence as input to downstream verification reduces fact checking accuracy by 17.2%. This indicates the necessity of retrieving evidential documents as top relevant documents for downstream claim verification. As justified by the example in Figure 1, by considering the verification success factor in the relevance assessment of documents, like the right list of documents in Figure 1, the updated top-ranked documents can include cleaner and richer information as evidence. However, the investigation of integrating such a verification success factor into evidence retrieval remains unexplored in the existing literature. Hence, in this study, we aim to advance the existing scientific fact checking system by improving the document retrieval component with a joint estimation of semantic relevance and downstream verification success of documents.

In particular, to validate the assumption of performance gain by including the verification success factor, we conduct a preliminary study to investigate the actual benefit of updating semantic relevance with additional verification feedback. A corresponding ideal document reranking scorer, ComboScorer, is introduced in this study. In addition, based on the concluding remarks of successful relevance integration, we further introduce a learned document reranking model, +VeriRel, that approximates the ComboScorer model without the additional cost of running verifiers. We conduct extensive experiments on three publicly available datasets, SciFact, SciFact-Open and CheckCOVID. The corresponding results consistently show the improvement of our approaches to both retrieval effectiveness and validation accuracy. The main contributions are as follows:

- We propose ComboScorer and +VeriRel to evaluate and leverage verification feedback in improving document ranking. The results show their consistent leading performance, compared to state-of-the-art baselines across three dataset.
- We explore the factors that can affect the scalability and domain generalisability of +VeriRel, which suggests a novel way to train effective document reranking models for the scientific fact checking task.

2 Related Work

2.1 Evidence retrieval

Existing scientific fact checking systems rely on semantic relevance ranking techniques for evidence retrieval. Many approaches rely on off-the-shelf approaches such as simple TF-IDF ranking or BM25 (Wadden et al., 2020; Wang et al., 2023). Learned models have also been applied. For example, Zhang et al. (2021); Li et al. (2021) employ BioSentVec (Chen et al., 2019), a biomedical version of Sent2Vec (Pagliardini et al., 2018), to compute similarity. Leveraging neural rerankers trained on general benchmark information retrieval datasets, e.g. MS MARCO (Bajaj et al., 2016), has also been shown to further improve evidence retrieval performance (Pradeep et al., 2021; Wadden et al., 2022b,a; Wührl and Klinger, 2021). Evidence retrieval may also be carried out in multiple stages, such as document retrieval followed by sentence selection (Vlachos and Riedel, 2014; Zeng et al., 2021; Guo et al., 2022).

2.2 Verification for Retrieval

Typical fact checking systems consist of separate components including for document retrieval, sentence selection and verification (Vlachos and Riedel, 2014). To mitigate error propagation and improve system integrity, these individual components are becoming increasingly interconnected, e.g. using joint learning (Zhang et al., 2021). Methods to leverage verification feedback to improve sentence selection have proved effective. For example, FER (Zhang et al., 2023) devise tailored loss functions to leverage the verification output difference between gold evidence and the feedback on selected sentences to improve the 'utility' of a sentence selection component. REREAD (Hu et al., 2023) also explores verification output to improve retrieval interpretability by introducing evidence metrics about sufficiency, fullness and plausibility. Similarly, Zheng et al. (2024) proposed a joint optimisation for retriever and reranker models learned from verification output to support sentence-as-evidence retrieval.

Although this related work demonstrates that sentence retrieval can be improved using feedback from downstream verification, the quality and generalizability of the claim verifiers used has a direct impact on the performance of this type of retriever. Indeed, biased verification models, caused by data imbalance (Torabi Asr and Taboada, 2019) and the negative sampling strategies (Wadden et al., 2022b; Vladika and Matthes, 2023) employed, can produce domain-specific retrievers with limited generalizability. This study develops verifiers that avoid this limitation. In addition, existing methods have

primarily focused on sentence selection while this work targets the improvement of document retrieval effectiveness which is a necessary pre-requisite for sentence-level fact checking.

3 Methodology

3.1 Problem Statement

This study focuses on the scientific fact checking that using top-retrieved documents to verify certain claims, which can be formulated as follows:

$$FC(c, D) \rightarrow \forall\, V(c, d),\ d \in D' \qquad (1)$$

where $FC(\cdot)$ refers to a scientific fact checking system with a claim, c, and document corpus, D, as input. FC can be further divided into two stages, evidence retrieval and verification. The evidence retrieval component uses the claim as a query, searches relevant documents from the given corpus and returns a smaller document subset (D') with a cutoff k:

$$D' = Retrieval(c, D, k) \qquad (2)$$

Retrieved documents within D' are then used for the subsequent verification stage (i.e., $V(\cdot)$):

$$V(c, d) = argmax(p^r_{c,d}, p^n_{c,d}, p^s_{c,d}) \qquad (3)$$

where $d \in D'$ and $p^r_{c,d}$, $p^s_{c,d}$ or $p^n_{c,d}$ are the estimates that d supports, refutes or does not provide enough information relative to c. The verifier computes a label based on the maximum value of the three probability scores. For example, if $p^s_{c,d} = 0.7$, $p^r_{c,d} = 0.2$ and $p^n_{c,d} = 0.1$, we obtain the label of will be "Support". After that, a scientific fact checking system will accumulate a set of claim/document pairs with corresponding the labels with respect to the three classes (i.e., support, refute or not enough information) to verify the claim c.

It seems intuitive that the use of relevant documents can have a direct impact on claim verification accuracy due to the availability of evidential information. Therefore this study focuses on the improvement of the document retrieval component to obtain useful documents for scientific fact checking. Note that we define the usefulness of documents based on whether they include evidential information for claim verification.

3.2 Integrating Verification

A typical document retrieval process calculates query-document relevance in one or two steps, initial retrieval and additional document reranking for the trade-off of effectiveness and efficiency (Hambarde and Proenca, 2023). The initial retrieval aims to include relevant documents in a pool with a high recall. After that, the document reranking stage focuses on identifying top-relevant documents from the document pool. As a consequence, by relying on many existing document retrieval pipelines, scientific fact checking systems can achieve a reasonable performance in obtaining semantically relevant documents. However, as argued in Section 1, fully relying on the semantic relevance to evaluate the usefulness of documents is insufficient for the scientific fact checking task and likely to introduce unwanted noise into the verification process. Meanwhile, a common fact checking pipeline includes evidence retrieval and verification components (see Section 3.1). The verification stage calculates the estimated likelihood of documents supporting, refuting claims or not having enough information. Hence, we assume that considering verification feedback in estimating claim-document relevance will improve retrieval effectiveness for scientific fact checking. This assumption is tested by developing an ideal document reranker.

3.2.1 ComboScorer

We devise ComboScorer, an ideal document reranker that combines both semantic relevance and feedback from a verifier to calculate the score for the final document ranking. Recall the definition of the verification method (see Eq. 3), we rely on the estimated probability of a document d supporting, refuting a claim c or not having enough information (i.e., $p^s_{c,d}$, $p^r_{c,d}$ or $p^n_{c,d}$) to justify the verification usefulness of such document. In particular, to effectively conduct downstream verification, we prefer documents that provide strong feedback to support or refute a claim, rather than not having enough information to support either decision. Hence, we integrate the support and refute likelihoods of documents to estimate their verification usefulness as follows:

$$s^v_{c,d} = p^s_{c,d} + p^r_{c,d} \qquad (4)$$

After that, similar to common document retrieval approaches, we still leverage the value of the semantic relevance between claim-document pairs to complete the relevance assessment. To be specific, by using an existing document reranking model, we can estimate the semantic relevance between claims and documents (i.e., $s^r_{c,d}$). Next, we calculate the final scores (i.e., s^{combo}) to rank documents

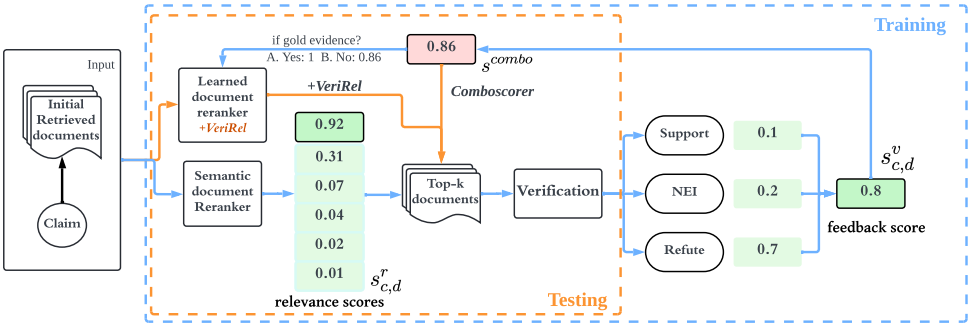

Figure 2: Pipeline to leverage verification feedback to enhance document retrieval. The blue area presents the flow of producing joint relevance scores for preliminary study and training +*VeriRel* reranker. The orange area shows the test of the ideal *ComboScorer* approach and trained +*VeriRel*.

by adopting a simple linear combination of the pre-calculated verification usefulness with the semantic relevance as follows:

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

where α controls the contribution of the semantic relevance and verification contributions. Hence, by integrating the semantic and verifiable features, ComboScorer is approaching an ideal solution that comprehensively assesses document relevance for scientific fact checking. The evaluation of ComboScorer can validate our preliminary assumption that the joint consideration of semantic relevance and verification success factor can effectively identify top useful documents for scientific fact checking. However, in a practical scenario, applying ComboScorer can be costly with repeated verification calculations. Hence, we further propose to model the relevance between claims and documents with a learned reranker that encapsulates both relevance features, semantics and verifiability.

3.3 +VeriRel

By considering the costly application of ComboScorer for document reranker, we propose a novel document reranking model, +VeriRel, which learns from the joint relevance and automatically estimates the usefulness of documents to verify a certain claim. The construction of +VeriRel can be formulated as follows:

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

where $f(c,d)$ is a function that models the relationship between a claim c and a document d. After applying the softmax function to constraint to the range of [0,1], so as to approximate the ideal ComboScorer (i.e., $s^{combo}_{c,d}$). Regarding the implementation of $f(c,d)$, in this study, we employ SciBERT (Beltagy et al., 2019), a pre-trained language model specifically designed for scientific literature and has shown to be effective in scientific applications (Wadden et al., 2020; Kotonya and Toni, 2020; Tan et al., 2023). Note that, to further improve the reliability of using $s^{combo}_{c,d}$, we update the $s^{combo}_{c,d}$ if the document d is labelled as the evidence in the training data.

Figure 2 presents an overview of the scientific fact checking pipeline used. Similar to many existing works, it consists of document initial and reranked retrieval, followed by a verification component. However, uniquely, by leveraging the feedback from verification, we first built *ComboScorer* to conduct the preliminary study. After that, we leverage the scores to further train our +*VeriRel* model to address the updated reranking schemes.

4 Experiments

In this section, we discuss the details of a series of experiments that we adopted 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.

4.1 Datasets

Three publicly available scientific fact checking datasets were used for the experiments: **SciFact** (Wadden et al., 2020) consists of 809 and 300 claims for train set and validate set, with a corpus of 5,183 high-quality scientific abstracts extracted from the S2ORC (Lo et al., 2020), a publicly-available corpus of millions of scientific articles. The test set contains 300 claims but the corresponding gold evidence is not publicly accessible. On the other hand, **SciFact-Open** (Wadden et al., 2022a), an extension of SciFact, having its test set using 279

claims from the test set of SciFact, while expanding the document corpus from 5,183 to 500,000 abstracts with additional evidence. Based on the overlap between SciFact and Scifact-Open, we update the test set of SciFact to the common 279 claims and use the corresponding label from SciFact-Open for evaluation. In addition, to avoid the repeated study on identical claim evidence, we follow the practice in Wadden et al. (2022a) of creating an independent testing dataset for SciFact-Open by ignoring the abstracts included in the SciFact dataset for separated evidence.

Figure 3 illustrates the updated SciFact and SciFact-Open datasets regarding the use of processed document corpus. The resulting SciFact and SciFact-Open test sets both include 279 claims but use different corpora are sources of evidence against which the claims can be verified. Note that in all the retrieval tasks in this study we consistently search from the full 500,000 documents to approximate the scenario of verifying against reasonably sized data sets.

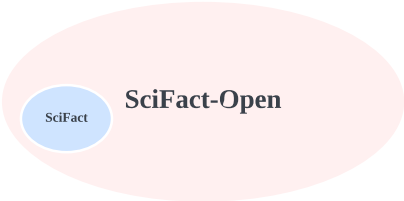

Figure 3: Illustration of relationships between the updated test sets of SciFact and SciFact-Open.

In addition to the SciFact and SciFact-Open dataset, we further include the third dataset, **Check-COVID** (Wang et al., 2023) consists of 1504 claims and a corpus including 347 scientific literature about COVID-19. Each claim has only one corresponding evidential document.

4.2 Semantic similarity-based retriever

Among the existing research works on SciFact and SciFact-Open, the combination of BM25 (Robertson et al., 1995) and T5-3B (Raffel et al., 2020) is the best-performing document retrieval pipeline and widely used (Pradeep et al., 2021; Wührl and Klinger, 2021; Wadden et al., 2022b,a). Hence, in our study, we adopt this retriever pipeline and set cut-off k in Eq. (2) as 500 for initial retrieval by BM25 on SciFact, 2000 on SciFact-Open, and 347 on Check-COVID for following reranking. T5-3B trained on MS MARCO and MS MARCO MED (i.e., a medical subset of MS MARCO), respectively, as the baseline reranking models. The use of the MS MARCO MED as an additional variant is due to the equal use of medical instances within the three datasets explored in this study.

4.3 Claim Verification

MultiVerS (Wadden et al., 2022b) is used as the verifier since it is the current state-of-the-art claim verifier on SciFact and SciFact-Open datasets according to the task leaderboard[1] and published results (Wadden et al., 2022b,a). MultiVerS uses the Longformer model (Beltagy et al., 2020) which enables processing of long documents and avoids information loss. MultiVerS is initialised with the checkpoint (Wadden et al., 2022b) trained on three datasets: FEVER (Thorne et al., 2018), PubMedQA (Jin et al., 2019) and Evidence Inference (Lehman et al., 2019; DeYoung et al., 2020)).

4.4 ComboScorer and +VeriRel Configurations

ComboScorer consists of (1) a semantic relevance score ($p^s_{c,d}$), which is calculated using a sigmoid function over the output of T5-3B, and (2) a verification feedback score ($p^r_{c,d}$) from MultiVerS by adding a softmax function to the output logits of three labels with its classification nature. To control the weight of two scores, the value of α in Eq. 5 was varied between 0 and 1 in steps of 0.1. The best performing result varies across negative sampling strategies but always performs best around 0.5 so this value is used for all experiments.

For the **+VeriRel** model, as mentioned in Section 3.3, SciBERT is to model the relationship between claim and documents to calculate a joint relevance score in the range of [0,1]. A score closer to 1 represents the most relevant to the claim. During the model training, we update the s^{combo} as semi-supervised labels to 1 if a document d is the gold evidence to a claim c and remain otherwise. For the reliability of the s^{combo} scores as ground truth, model training data is limited to the top 20 documents ranked by ComboScorer. For the evidence documents not included in top 20 documents, we added them to the training data with their label set to 1. In addition, we train the model using a learning rate of 1e-5 with Adam optimizer and apply the get_cosine_schedule_with_warmup setup provided by transformer with 40 epochs.

To investigate the verifier's generalisability, Mul-

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

Model	SciFact		
	Precision	Recall	F1
V-MultiVerS(N20)	**62.16**	72.52	**66.94**
V-MultiVerS(N10)	57.71	72.52	64.27
V-MultiVerS(N5)	41.45	**77.48**	54.00

Table 1: Verification performance on SciFact.

tiVerS is trained on the SciFact dataset with varied sizes of negative samples (i.e., $\{5, 10, 20\}$) obtained by randomly sampling from the top 100 documents ranked by BM25. The resulting verifier, which is used to provide verification feedback, is named 'V-MultiVerS' to distinguish it from the off-the-shelf 'MultiVerS' used to examine verification improvements in later experiments. Table 1 presents this corresponding effect of varying negative sampling strategies on verification performance with results obtained by submitting to the task leaderboard. The results show that using a larger number of negative samples (i.e., N(20)) during verifier training improves the precision (i.e., higher specificity) while fewer negative samples (i.e., N(5)) improves the generalizability with higher recall. (All model training was performed with the same random seed, 27, except the reproducibility study, using a single NVIDIA A100 GPU. We will release the code for reproducibility upon paper acceptance.)

4.5 Evaluation

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

$$Recall@k = \frac{N(relevant@k)}{N(relevant)} \quad (7)$$

where $N(relevant@k)$ and $N(relevant)$ are the number of relevant documents in the top k and entire corpus.

Hit metrics provide additional information about document retrieval effectiveness. **hit-one**: The proportion of claims for which at least one of the relevant documents is included in those retrieved. **hit-all**: The proportion for which all relevant documents are included. Since no evidence (i.e. relevant documents) is available for some claims, two variants of these metrics are introduced, **hit-one evi** and **hit-all evi**, which only consider the claims for which evidence is available.

Verification performance is measured using the standard metrics for the task (precision, recall and F1) which are computed via the automatic evaluation for SciFact in the leaderboard provided for the SCIVER shared task (Wadden and Lo, 2021), using provided 10 retrieved documents.

5 Results

This section presents the performance evaluation of the proposed ComboScorer method and +VeriRel models. We compare the retrieval effectiveness of our approaches to baselines, emphasizing the improvements in document ranking and verification success achieved through the integration of verification feedback. The experimental results show that both ComboScorer and +VeriRel outperform baseline methods across multiple datasets, demonstrating their generalisable leading performance.

5.1 ComboScorer

We first validate the ideal effect of combining verification feedback into the claim-document relevance by evaluating *ComboScorer*. Table 2 presents the performance of ComboScorer by leveraging verification feedback as shown in Figure 2. ComboScore performance for all three evaluation corpora is shown in the middle rows below the baselines. Results show that ComboScorer consistently improves document retrieval by retrieving more evidence across the settings of various evaluation cutoffs (i.e., k = 1 to 50). In particular, wider improvement margins are observed in the top positions with smaller values of cut-off k (k<10). For example, when setting k to 3, regarding the best-performing choice of negative sample size, we observe that ComboScorer(N20) on SciFact can outperform the top baseline with around 10% improvements. For another two datasets, SciFact-Open and Check-COVID, around 20% and 5% of improvements can be observed when using ComboScorer trained on 5 negative samples.

Recall the discussion in Section 4.4, we develop verifiers to provide verification feedback for ComboScorer by training the verfier on the SciFact dataset. This enables the investigation of in-domain and out-of-the-domain evaluations. By looking into the experimental results in Table 1, we observe that ComboScorer trained with a V-MultiVerS and setting the number of negative samples to 20 can enable best in-domain performance (i.e., its leading performance on SciFact). On another two datasets about out-of-domain scenarios, the use of V-MultiVerS trained on few negative samples (N=5) consistently results in the best performance, which indicates the value of a general verifier for

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
T5-3B(MS MARCO)	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
T5-3B(MS MARCO 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
ComboScorer(N20)	**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
ComboScorer(N10)	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
ComboScorer(N5)	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**
+VeriRel(baseline)	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(N20)	**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
+VeriRel(N10)	**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
+VeriRel(N5)	**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 2: Performance of baselines, ComboScorer and +VeriRel in Recall@k with cutoff k ranges from 50 to 1.

out-of-domain evidence retrieval.

Next, we investigate the impact of ComboScorer with respect to the ratio of retrieving relevant documents to claims via the hit metrics discussed in Section 4.5. Table 3 presents the experimental results evaluated by the hit metrics on the SciFact and SciFact-Open datasets. Check-Covid dataset is not included since it has only one relevant evidential document to each claim and the corresponding performance can be indicated via Recall@k in Table 1. According to the experimental results, we observe that ComboScorer can consistently outperform baselines to retrieve more relevant documents for all claims on average with higher hit-all and hit-one scores. Again, we observe the identical in-domain and out-of-domain effects while using verifiers that are trained on different numbers of negative samples (i.e., setting N to 20 for in-domain and 5 for out-of-domain scenarios).

Overall, by validating our assumption in this preliminary study, we conclude that including verification feedback to assess document relevance, in addition to semantic relevance, can indeed improve the retrieval effectiveness to identify relevant evidential documents for scientific fact checking.

Method/Top 20	SciFact			
	hit-one evi	hit-all evi	hit-one	hit-all
BM25	71.58	69.47	80.65	79.21
T5-3B(MS MARCO)	90.00	87.89	93.19	91.76
T5-3B(MS MARCO MED)	90.00	87.89	93.19	91.76
Comboscorer(N20)	**92.63**	**90.00**	**94.98**	**93.19**
Comboscorer(N10)	92.11	89.47	94.62	92.83
Comboscorer(N5)	91.58	88.95	94.27	92.47
Method/Top 20	SciFact-Open			
	hit-one evi	hit-all evi	hit-one	hit-all
BM25	76.54	38.27	93.19	82.08
T5-3B(MS MARCO)	96.30	61.73	98.92	88.89
T5(MS MARCO MED)	95.06	59.26	98.57	88.17
Comboscorer(N20)	98.77	66.67	99.64	90.32
Comboscorer(N10)	98.77	66.67	99.64	90.32
Comboscorer(N5)	**100**	**71.60**	**100**	**91.76**

Table 3: Reported hit metrics on SciFact and SciFact-Open.

5.2 +VeriRel performance

After validating the assumption about the value of adding verification feedback to retrieval, we turn to the evaluation of our +VeriRel model, a trained ranker that approximates ComboScorer. Table 1 includes the experimental results on three datasets. According to the results, +VeriRel outperforms the baselines under most evaluation circumstances. Performance for Recall @10 and @3 on the SciFact-Open dataset are exceptions but +VeriRel's performance is still close to the most competitive baseline. To ensure a fair comparison, we also include +VeriRel trained on the semantic relevance score only, calculated by the T5 model and named +VeriRel(baseline). The experimental results indicate the effective approximation to ComboScorer with advanced retrieval performance. Specifically, when comparing the performance between +VeriRel and ComboScorer, we observe that ComboScorer as an ideal scoring function can still outperform +VeriRel in most cases apart from the evaluation on the Check-COVID dataset with the evaluation cutoff larger than 1. This can be caused by the limitation of the used SciBERT model to process long documents, since 27.4% of inputs, combining claim and abstract, exceed 512 tokens, which is the token limit of SciBERT. This could result in information loss to allow effective relevance assessment. Hence, we aim to further improve the +VeriRel with more advanced language processing models as backbones to address long-context inputs in future studies.

5.3 Verification improvement with +VeriRel

After validating the successful retrieval improvement with our proposed +VeriRel model, it is essential to evaluate whether the improvement can further benefit the downstream verification effectiveness. Table 4 shows the verification accuracy based on using +VeriRel and the state-of-the-art T5-3B(MS MARCO) model. To fairly compare

performance, we use MultiVerS, the state-of-the-art verifier on SciFact, and initiate using the released checkpoint (Wadden et al., 2022b) for verification assessment. Note that, the experimental results on SciFact-Open and Check-COVID are not included due to the limited generalizability of MultiVerS. For example, it predicts 204 oracle evidence out of 251 as 'not enough information' on the SciFact-Open dataset, making it difficult to conclude meaningful insights. For the retrieval setup, we first set the initial retrieval cutoff to 2,000 and then rerank the documents to get the top 3, 5 and 10 documents for downstream verification. We refer to this evaluation as SciFact (offline). We also run an additional test setup by submitting the results to the public SciFact leaderboard provided by the shared task SCIVER (Wadden et al., 2020; Wadden and Lo, 2021). We follow Wadden et al. (2022b) by using the top 20 initial retrieved documents, instead of the top 2,000 documents, by BM25 and then rerank to get the top documents.

Model	Verification performance			Retrieval performance
+ MultiVerS	P	R	F1	Recall@k
Top 10	SciFact(offline)			Recall@10
T5-3B(MS MARCO)	74.52	55.98	63.93	86.60
+VeriRel(N5)	**75.88**	**61.72**	**68.07**	**88.04**
Top 5	SciFact(offline)			Recall@5
T5-3B(MS MARCO)	**73.03**	62.20	67.18	80.38
+VeriRel(N5)	72.48	**65.55**	**68.84**	**84.21**
Top 3	SciFact(offline)			Recall@3
T5-3B(MS MARCO)	76.43	57.42	65.57	73.68
+VeriRel(N5)	**77.78**	**63.64**	**70.00**	**78.95**
Top 10	SciFact(leaderboard)			
T5-3B(MS MARCO)	**73.83**	**71.17**	**72.48**	Not accessible
+VeriRel(N5)	**73.83**	**71.17**	**72.48**	Not accessible
Top 5	SciFact(leaderboard)			
T5-3B(MS MARCO)	74.88	69.82	72.26	Not accessible
+VeriRel(N5)	**75.12**	**70.72**	**72.85**	Not accessible
Top 3	SciFact(leaderboard)			
T5-3B(MS MARCO)	**77.04**	68.08	72.25	Not accessible
+VeriRel(N5)	75.86	**69.37**	**72.47**	Not accessible

Table 4: Verification performance by inputting document retrieved by proposed +VeriRel and baseline.

Regarding the evaluation setup of SciFact(offline), +VeriRel can consistently improve the verification accuracy with higher F1 scores, when compared to the semantic relevance-based approach. The verification performance can achieve a 70% F1 score with the choice of using the top 3 relevant documents for verification. Meanwhile, by comparing the experimental results when submitting to the leaderboard, we observe that +VeriRel can still consistently improve the baseline for improved verification performance with the maximum 72.85% F1 score.

In addition, we also list the top-performing verification approaches in the leaderboard in Table 5 and show that our solution can advance the best-performing MultiVerS with a higher F1 score with a wider margin than the second-ranked approach (i.e., ARSJOINT). These findings indicate that our proposed +VeriRel can improve the retrieval effectiveness, which consistently benefits the downstream verification performance.

Model	Verification performance		
	P	R	F1
VerT5erini	64.03	72.97	68.21
ParagraphJoint	**75.81**	63.51	69.12
MultiVerS	73.83	**71.17**	72.48
ARSJOINT	72.22	70.27	71.23
MultiVerS +VeriRel(N5)	75.12	70.72	**72.85**

Table 5: Top fact checking systems in the leaderboard.

6 Conclusion

This study presents +VeriRel, a novel approach that enhances document retrieval for scientific fact checking by leveraging feedback from verification stages. By incorporating a verification reward model into the ranking mechanism, +VeriRel consistently improves retrieval accuracy, ensuring that more relevant and evidential documents are prioritized. Our experiments show that +VeriRel outperforms traditional methods in both scalability and generalization, especially when dealing with large, diverse, and previously unseen corpora. 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. In particular, according to our findings, we encourage the separation of the training of the verification reward model and the claim verifier. Initially, the verification reward model should be trained with a smaller number of negative samples, allowing generalisable identification of relevant evidence. Once the reranker is optimized using this feedback, a separate, tailored claim verifier can be trained for the inference stage.

Our findings highlight that the key to improving document retrieval lies in the careful balance of feedback integration, with fewer negative samples offering better scalability and adaptability to new datasets. This novel approach not only bridges a critical gap in the existing fact checking pipeline but also paves the way for future enhancements in scientific fact checking systems by effectively linking retrieval and verification components.

Limitation

While VeriFED shows promise in improving document retrieval, there are several limitations to consider. A limitation is the truncation of input data when using models like SciBERT, which affects the quality of the reranking process due to the loss of critical information beyond the token limit. Moreover, the feedback mechanism relies heavily on the quality of the verification model, which means that inaccuracies in the verifier can propagate and affect the retrieval performance. Addressing these limitations will require further research into optimizing model efficiency, reducing biases, and improving the scalability of the verification-feedback mechanism.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

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)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

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

Tuhin Chakrabarty, Tariq Alhindi, and Smaranda Muresan. 2018. Robust document retrieval and individual evidence modeling for fact extraction and verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 127–131, Brussels, Belgium. Association for Computational Linguistics.

Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2019. Biosentvec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5. IEEE.

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*, pages 123–132, Online. Association for Computational Linguistics.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Kailash A Hambarde and Hugo Proenca. 2023. Information retrieval: recent advances and beyond. *IEEE Access*.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

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*, pages 2319–2323.

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)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

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)*, pages 7740–7754, Online. Association for Computational Linguistics.

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)*, pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiangci Li, Gully A Burns, and Nanyun Peng. 2021. A paragraph-level multi-task learning model for scientific fact-verification. In *SDU@ AAAI*.

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.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *Proceedings of the*

2023 Conference on Empirical Methods in Natural Language Processing, pages 7787–7813, Singapore. Association for Computational Linguistics.

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*, pages 244–257, Marseille, France. European Language Resources Association.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. 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 528–540, New Orleans, Louisiana. Association for Computational Linguistics.

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*, pages 94–103, online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

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.

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*, pages 1785–1789.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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*, pages 2652–2664, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Fatemeh Torabi Asr and Maite Taboada. 2019. Big data and quality data for fake news and misinformation detection. *Big data & society*, 6(1):2053951719843310.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.

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*, pages 6215–6230, Toronto, Canada. Association for Computational Linguistics.

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 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*, pages 124–129, Online. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022a. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022b. 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.

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*, pages 14114–14127, Toronto, Canada. Association for Computational Linguistics.

Amelie Wührl and Roman Klinger. 2021. Claim detection in biomedical Twitter posts. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 131–142, Online. Association for Computational Linguistics.

Xia Zeng, Amani S Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10):e12438.

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*, pages 6373–6384, Singapore. Association for Computational Linguistics.

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*, pages 3580–3586, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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 and virtual meeting. Association for Computational Linguistics.

Appendix

A Hit metrics of ComboScorer

Method/Top 50	SciFact			
	hit-one evi	hit-all evi	hit-one	hit-all
BM25	77.37	74.74	84.59	82.80
T5-3B(MS MARCO)	93.16	91.05	95.34	93.91
T5-3B(MS MARCO MED)	93.68	91.58	95.70	94.27
ComboScorer(N20)	**95.26**	**92.63**	**96.77**	**94.98**
ComboScorer(N10)	94.74	92.11	96.42	94.62
ComboScorer(N5)	94.74	92.11	96.42	94.62
Method/Top 20	SciFact			
	hit-one evi	hit-all evi	hit-one	hit-all
BM25	71.58	69.47	80.65	79.21
T5-3B(MS MARCO)	90.00	87.89	93.19	91.76
T5-3B(MS MARCO MED)	90.00	87.89	93.19	91.76
ComboScorer(N20)	**92.63**	**90.00**	**94.98**	**93.19**
ComboScorer(N10)	92.11	89.47	94.62	92.83
ComboScorer(N5)	91.58	88.95	94.27	92.47
Method/Top 10	SciFact			
	hit-one evi	hit-all evi	hit-one	hit-all
BM25	65.79	62.63	76.70	74.55
T5-3B(MS MARCO)	88.95	87.37	92.47	91.40
T5-3B(MS MARCO MED)	88.42	86.32	92.11	90.68
ComboScorer(N20)	89.47	**87.89**	92.83	**91.76**
ComboScorer(N10)	**90.53**	**87.89**	**93.55**	**91.76**
ComboScorer(N5)	88.95	86.84	92.47	91.04
Method/Top 5	SciFact			
	hit-one evi	hit-all evi	hit-one	hit-all
BM25	59.47	56.84	72.40	70.61
T5-3B(MS MARCO)	82.63	80.53	88.17	86.74
T5-3B(MS MARCO MED)	83.16	81.05	88.53	87.10
ComboScorer(N20)	84.21	82.63	89.25	88.17
ComboScorer(N10)	**85.79**	**84.21**	**90.32**	**89.25**
ComboScorer(N5)	85.79	82.63	90.32	88.17
Method/Top 3	SciFact			
	hit-one evi	hit-all evi	hit-one	hit-all
BM25	52.63	50.53	67.74	66.31
T5-3B(MS MARCO)	75.79	72.63	83.51	81.36
T5-3B(MS MARCO MED)	76.84	74.21	84.95	82.44
ComboScorer(N20)	**82.11**	**80.53**	**87.81**	**86.74**
ComboScorer(N10)	82.11	78.95	87.81	85.66
ComboScorer(N5)	78.95	75.79	85.66	83.51
Method/Top 1	SciFact			
	hit-one evi	hit-all evi	hit-one	hit-all
BM25	38.95	37.89	58.42	57.71
T5-3B(MS MARCO)	60.53	57.89	73.12	71.33
T5-3B(MS MARCO MED)	61.05	58.42	73.48	71.68
ComboScorer(N20)	**67.89**	**65.26**	**78.14**	**76.34**
ComboScorer(N10)	66.32	63.16	77.06	74.91
ComboScorer(N5)	61.58	58.95	73.84	72.04

Table 6: Top-k hit metrics / SciFact

Method/Top 50	SciFact-Open			
	hit-one evi	hit-all evi	hit-one	hit-all
BM25	88.89	51.85	96.77	86.02
T5-3B(MS MARCO)	98.77	76.54	99.64	93.19
T5-3B(MS MARCO MED)	98.77	75.31	99.64	92.83
ComboScorer(N20)	**100**	81.48	**100**	94.62
ComboScorer(N10)	**100**	82.72	**100**	94.98
ComboScorer(N5)	**100**	**83.95**	**100**	**95.34**
Method/Top 20	SciFact-Open			
	hit-one evi	hit-all evi	hit-one	hit-all
BM25	76.54	38.27	93.19	82.08
T5-3B(MS MARCO)	96.30	61.73	98.92	88.89
T5-3B(MS MARCO MED)	95.06	59.26	98.57	88.17
ComboScorer(N20)	98.77	66.67	99.64	90.32
ComboScorer(N10)	98.77	66.67	99.64	90.32
ComboScorer(N5)	**100**	**71.60**	**100**	**91.76**
Method/Top 10	SciFact-Open			
	hit-one evi	hit-all evi	hit-one	hit-all
BM25	59.26	28.40	88.17	79.21
T5-3B(MS MARCO)	88.89	54.32	96.77	86.74
T5-3B(MS MARCO MED)	87.65	53.09	96.42	86.38
ComboScorer(N20)	96.30	**61.73**	98.92	**88.89**
ComboScorer(N10)	**97.53**	58.02	**99.28**	87.81
ComboScorer(N5)	92.59	53.09	97.85	86.38
Method/Top 5	SciFact-Open			
	hit-one evi	hit-all evi	hit-one	hit-all
BM25	49.38	20.99	85.30	77.06
T5-3B(MS MARCO)	77.78	38.27	93.55	82.08
T5-3B(MS MARCO MED)	79.01	40.74	93.91	82.80
ComboScorer(N20)	85.19	39.51	95.70	82.44
ComboScorer(N10)	85.19	38.27	95.70	82.08
ComboScorer(N5)	**87.65**	**44.44**	**96.42**	**83.87**
Method/Top 3	SciFact-Open			
	hit-one evi	hit-all evi	hit-one	hit-all
BM25	40.74	18.52	82.80	76.34
T5-3B(MS MARCO)	60.49	25.93	88.53	78.49
T5-3B(MS MARCO MED)	61.73	28.40	88.89	79.21
ComboScorer(N20)	71.60	27.16	91.76	78.85
ComboScorer(N10)	76.54	30.86	93.19	79.93
ComboScorer(N5)	**77.78**	**32.10**	**93.55**	**80.29**
Method/Top 1	SciFact-Open			
	hit-one evi	hit-all evi	hit-one	hit-all
BM25	23.46	6.17	77.78	72.76
T5-3B(MS MARCO)	34.57	7.41	81.00	73.12
T5-3B(MS MARCO MED)	32.10	6.17	80.29	72.76
ComboScorer(N20)	35.80	8.64	81.36	73.48
ComboScorer(N10)	34.57	8.64	81.00	73.48
ComboScorer(N5)	**37.04**	**9.88**	**81.72**	**73.84**

Table 7: Top-k hit metrics / SciFact-Open