SCIFACT-OPEN: Towards open-domain scientific claim verification

David Wadden[†]   **Kyle Lo**[‡]   **Bailey Kuehl**[‡]   **Arman Cohan**[‡]
Iz Beltagy[‡]   **Lucy Lu Wang**[†‡]   **Hannaneh Hajishirzi**[†‡]
[†] University of Washington, Seattle, WA, USA
[‡] Allen Institute for Artificial Intelligence, Seattle, WA, USA
{dwadden,hannaneh}@cs.washington.edu, lucylw@uw.edu,
{kylel,baileyk,armanc,beltagy}@allenai.org

Abstract

While research on scientific claim verification has led to the development of powerful systems that appear to approach human performance, these approaches have yet to be tested in a realistic setting against large corpora of scientific literature. Moving to this open-domain evaluation setting, however, poses unique challenges; in particular, it is infeasible to exhaustively annotate all evidence documents. In this work, we present SCIFACT-OPEN, a new test collection designed to evaluate the performance of scientific claim verification systems on a corpus of 500K research abstracts. Drawing upon pooling techniques from information retrieval, we collect evidence for scientific claims by pooling and annotating the top predictions of four state-of-the-art scientific claim verification models. We find that systems developed on smaller corpora struggle to generalize to SCIFACT-OPEN, exhibiting performance drops of at least 15 F1. In addition, analysis of the evidence in SCIFACT-OPEN reveals interesting phenomena likely to appear when claim verification systems are deployed in practice, e.g., cases where the evidence supports only a special case of the claim. Our dataset is available at https://github.com/dwadden/scifact-open.

1 Introduction

The task of scientific claim verification (Wadden et al., 2020; Kotonya and Toni, 2020) aims to help system users assess the veracity of a scientific claim relative to a corpus of research literature. Most existing work and available datasets focus on verifying claims against a much more limited context—for instance, a single article or text snippet (Saakyan et al., 2021; Sarrouti et al., 2021; Kotonya and Toni, 2020) or a small, artificially-constructed collection of documents (Wadden et al., 2020). Current state-of-the-art models are able to achieve very strong performance on these datasets, in some cases approaching human agreement (Wadden et al., 2022).
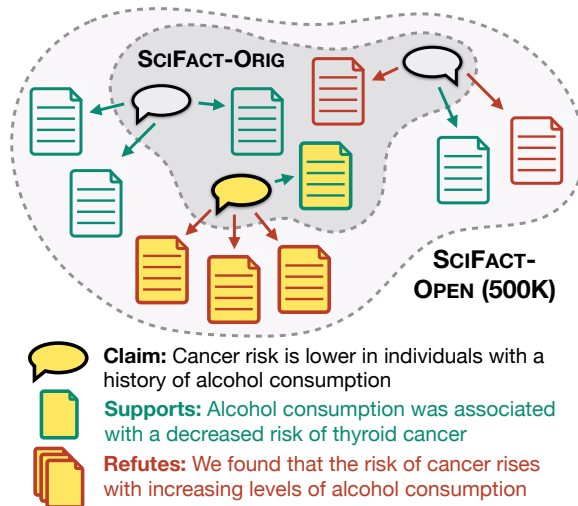

Figure 1: SCIFACT-OPEN, a new test collection for scientific claim verification that expands beyond the 5K abstract retrieval setting in the original SCIFACT dataset (Wadden et al., 2020) to a corpus of 500K abstracts. Each claim in SCIFACT-OPEN is annotated with evidence that SUPPORTS or REFUTES the claim. In the example shown, the majority of evidence REFUTES the claim that alcohol consumption reduces cancer risk, although one abstract indicates that alcohol consumption may reduce thyroid cancer risk specifically.

This gives rise to the question of the scalability of scientific claim verification systems to realistic, *open-domain* settings that involve verifying claims against corpora containing hundreds of thousands of documents. In these cases, claim verification systems should assist users by identifying and categorizing all available documents that contain evidence supporting or refuting each claim (Fig. 1). However, evaluating system performance in this setting is difficult because exhaustive evidence annotation is infeasible, an issue analogous to evaluation challenges in information retrieval (IR).

In this paper, we construct a new test collection for open-domain scientific claim verification, called SCIFACT-OPEN, which requires models to verify claims against evidence from both the SCIFACT (Wadden et al., 2020) collection, as well as

4748

additional evidence from a corpus of 500K scientific research abstracts. To avoid the burden of exhaustive annotation, we take inspiration from the *pooling* strategy (Sparck Jones and van Rijsbergen, 1975) popularized by the TREC competitions (Voorhees and Harman, 2005) and combine the predictions of several state-of-the-art scientific claim verification models—for each claim, abstracts that the models identify as likely to SUPPORT or REFUTE the claim are included as candidates for human annotation.

Our main contributions and findings are as follows. (1) We introduce SCIFACT-OPEN, a new test collection for open-domain scientific claim verification, including 279 claims verified against evidence retrieved from a corpus of 500K abstracts. (2) We find that state-of-the-art models developed for SCIFACT perform substantially worse (at least 15 F1) in the open-domain setting, highlighting the need to improve upon the generalization capabilities of existing systems. (3) We identify and characterize new dataset phenomena that are likely to occur in real-world claim verification settings. These include mismatches between the specificity of a claim and a piece of evidence, and the presence of conflicting evidence (Fig. 1).

With SCIFACT-OPEN, we introduce a challenging new test set for scientific claim verification that more closely approximates how the task might be performed in real-word settings. This dataset will allow for further study of claim-evidence phenomena and model generalizability as encountered in open-domain scientific claim verification.

2 Background and Task Overview

We review the scientific claim verification task, and summarize the data collection process and modeling approaches for SCIFACT, which we build upon in this work. We elect to use the SCIFACT dataset as our starting point because of the diversity of claims in the dataset and the availability of a number of state-of-the-art models that can be used for pooled data collection. In the following, we refer to the original SCIFACT dataset as SCIFACT-ORIG.

2.1 Task definition

Given a claim c and a corpus of research abstracts $\mathcal{A}$, the scientific claim verification task is to identify all abstracts in $\mathcal{A}$ which contain evidence relevant to c, and to predict a label $y(c, a) \in \{\text{SUPPORTS}, \text{REFUTES}\}$ for each evidence abstract. All other abstracts are labeled

$y(c, a) = \text{NEI}$ (Not Enough Info). We will refer to a single (c, a) pair as a *claim / abstract pair*, or CAP. Any CAP where the abstract a provides evidence for the claim c (either SUPPORTS or REFUTES) will be called an *evidentiary CAP*, or ECAP. Models are evaluated on their precision, recall, and F1 in identifying and correctly labeling the evidence abstracts associated with each claim in the dataset (or equivalently, in identifying ECAPs).[1]

2.2 SCIFACT-ORIG

Each claim in SCIFACT-ORIG was created by rewriting a citation sentence occurring in a scientific article, and verifying the claim against the abstracts of the cited articles. The resulting claims are diverse both in terms of their subject matter—ranging from molecular biology to public health—as well as their level of specificity (see §3.3). Models are required to retrieve and label evidence from a small (roughly 5K abstract) corpus.

Models for SCIFACT-ORIG generally follow a two-stage approach to verify a given claim. First, a small collection of candidate abstracts is retrieved from the corpus using a retrieval technique like BM25 (Robertson and Zaragoza, 2009); then, a transformer-based language model (Devlin et al., 2019; Raffel et al., 2020) is trained to predict whether each retrieved document SUPPORTS, REFUTES, or contains no relevant evidence (NEI) with respect to the claim.

As we show in §4 and §5, a key determinant of system generalization is the *negative sampling ratio*. A negative sampling ratio of r indicates that the model is trained on r irrelevant CAPs for every relevant ECAP. Negative sampling has been shown to improve performance (particularly precision) on SCIFACT-ORIG (Li et al., 2021). See Appendix A.4 for additional details.

3 The SCIFACT-OPEN dataset

In this section, we describe the construction of SCIFACT-OPEN. We report the performance of claim verification models on SCIFACT-OPEN in §4, and perform reliability checks on the results in §5.

Our goal is to construct a test collection which can be used to assess the performance of claim verification systems deployed on a large corpus

---

[1]The original SCIFACT task also requires the prediction of rationales justifying each label. Due to the expense of collecting rationale annotations, in this work we do not require rationales; we evaluate using the *abstract-level label-only F1* metric described in Wadden et al. (2020).
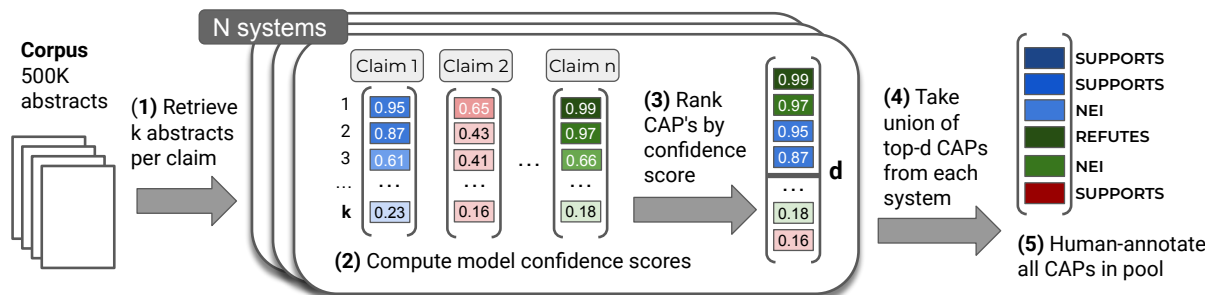
Figure 2: Pooling methodology used to collect evidence for SCIFACT-OPEN. We construct the pool by combining the d most-confident predictions of n different systems. A single CAP is represented as a colored box; the number in the box indicates a hypothetical confidence score. In this example, the annotation pool contains 3 CAPs from Claim 1, 2 for Claim 2, and 1 for Claim 3. Annotators found evidence for 4 / 6 of these CAPS.

of scientific literature. This requires a collection of claims, a corpus of abstracts against which to verify them, and evidence annotations with which to evaluate system predictions. We use the claims from the SCIFACT-ORIG test set as our claims for SCIFACT-OPEN.[2] To obtain evidence annotations, we use all evidence from SCIFACT-ORIG as evidence in our new dataset and collect additional evidence from the SCIFACT-OPEN corpus.

For our corpus, we filter the S2ORC dataset (Lo et al., 2020) for all articles which (1) cover topics related to medicine or biology and (2) have at least one inbound and one outbound citation. From the roughly 6.5 million articles that pass these filters, we randomly sample 500K articles to form the corpus for SCIFACT-OPEN, making sure to include the 5K abstracts from SCIFACT-ORIG. We choose to limit the corpus to 500K abstracts to ensure that we can achieve sufficient annotation coverage of the available evidence. Additional details on corpus construction can be found in Appendix A.

Unlike SCIFACT-ORIG (which is skewed toward highly-cited articles from "high-impact" journals), we do not impose any additional quality filters on articles included in SCIFACT-OPEN; thus, our corpus captures the full diversity of information likely to be encountered when scientific fact-checking systems are deployed on real-world resources like S2ORC, arXiv,[3] or PubMed Central.[4]

3.1 Pooling for evidence collection

To collect evidence from the SCIFACT-OPEN corpus, we adopt a pooling approach popularized by

the TREC competitions: use a collection of state-of-the-art models to select CAPs for human annotation, and assume that all un-annotated CAPs have $y(c, a) = \text{NEI}$. We will examine the degree to which this assumption holds in §5.

Pooling approach We annotate the d most-confident predicted CAPS from each of n claim verification systems. An overview of the process is in shown in Fig. 2; we number the annotation steps below to match the figure.

We select the most confident predictions for a single model as follows. **(1)** For each claim in SCIFACT-OPEN, we use an information retrieval system consisting of BM25 followed by a neural re-ranker (Pradeep et al., 2021) to retrieve k abstracts from the SCIFACT-OPEN corpus. **(2)** For each CAP, we compute the softmax scores associated with the three possible output labels, denoted $s(\text{SUPPORTS}), s(\text{REFUTES}), s(\text{NEI})$. We use $\max(s(\text{SUPPORTS}), s(\text{REFUTES}))$ as a measure of the model's confidence that the CAP contains evidence. **(3)** We rank all CAPs by model confidence, and add the d top-ranked predictions to the annotation pool. The final pool **(4)** is the union of the top-d CAPs identified by each system. Since some CAPs are identified by multiple systems, the size of the final annotation pool is less than $n \times d$; we provide statistics in §3.2. Finally, **(5)** all CAPs in the pool are annotated for evidence and assigned a final label by an expert annotator, and the label is double-checked by a second annotator (see Appendix A for details).

We choose to prioritize CAPS for annotation based on model confidence, rather than annotating a fixed number of CAPs per claim, in order to maximize the amount of evidence likely to be discovered during pooling. In §3.3, we confirm that our pro-

---

[2]We remove 21 claims (out of 300 total) whose source citations lack important metadata; see Appendix A for details.

[3]https://arxiv.org

[4]https://www.ncbi.nlm.nih.gov/pmc

Model	Source	Negative sampling
Pooling and Eval		
VERT5ERINI	Pradeep et al. (2021)	0
PARAGRAPHJOINT	Li et al. (2021)	10
MULTIVERS	Wadden et al. (2022)	20
MULTIVERS$_{10}$	Wadden et al. (2022)	10
Eval only		
ARSJOINT	Zhang et al. (2021)	12

Table 1: Models used for pooled data collection and evaluation (top), and for evaluation only (bottom). "Negative sampling" indicates the negative sampling ratio. MULTIVERS$_{10}$ shares the same architecture as MULTIVERS, but trains on fewer negative samples.

cedure identifies more evidence for claims that we would expect to be more extensively-studied.

Models and parameter settings We set $k = 50$ for abstract retrieval. In practice, we found that the great majority of evidentiary abstracts were ranked among the top 20 retrievals for their respective claims (Appendix A.3), and thus using a larger k would serve mainly to increase the number of irrelevant results. We set $d = 250$; in §5.1, we show that this is sufficient to ensure that our dataset can be used for reliable model evaluation.

For our models, we utilized all state-of-the-art models developed for SCIFACT-ORIG for which modeling code and checkpoints were available (to our knowledge). We used $n = 4$ systems for pooled data collection. During evaluation, we included a fifth system — ARSJOINT— which became available after the dataset had been collected. Model names, source publications, and negative sampling ratios are listed in Table 1; see Appendix A for additional details.

3.2 Dataset statistics

We summarize key properties of SCIFACT-OPEN. Table 2a provides an overview of the claims, corpus, and evidence in the dataset. Table 2b shows the fraction of CAPs annotated during pooling which were judged to be ECAPs (i.e. to contain evidence). Overall, roughly a third of predicted CAPs were judged as relevant; this indicates that existing systems achieve relatively low precision when used in an open-domain setting. Relevance is somewhat higher (roughly 50%) for CAPs predicted by more than one system. The majority of CAPs are selected by a single system only, indicating high diversity in model predictions. As mentioned in §3.1, the

		ECAPs		
Claims	Corpus	SCIFACT-ORIG	Pooling	Total
279	500K	209	251	460

(a) Summary of the SCIFACT-OPEN dataset, including the number of claims, abstracts, and ECAPs (evidentiary claim / evidence pairs). ECAPs come from two sources: those from SCIFACT-ORIG, and those discovered via pooling.

Num. systems	Annotated	Evidence	% Evidence
1	528	154	29.2
2	150	71	47.3
3	44	20	45.5
4	10	6	60.0
All	732	251	34.3

(b) Relevance of CAPs annotated during the pooling process. The first row indicates that 528 CAPs were identified for pooling by one system only; of those CAPs, 154 were judged by annotators as containing evidence. The more systems identified a given CAP, the more likely it is to contain evidence.

	Total	Retrieved	Annotated
ECAPs	209	187 (89%)	171 (82%)

(c) Count of how many ECAPs from SCIFACT-ORIG would have been identified during pooled data collection. "Retrieved" indicates the number of ECAPs that would have been retrieved among the top k, and "Annotated" indicates the number that would further have been included in the annotation pool.

Table 2: Annotation results and dataset statistics for SCIFACT-OPEN.

total number of annotated CAPs is 732 (rather than 4 models × 250 CAPs / model = 1000) due to overlap in system predictions.

Table 2c shows how many of the ECAPs from SCIFACT-ORIG would have been annotated by our pooling procedure. The fact that the great majority of the original ECAPs would have been included in the annotation pool suggests that our approach achieves reasonable evidence coverage.

3.3 Evidence phenomena in SCIFACT-OPEN

We observe three properties of evidence in SCIFACT-OPEN that have received less attention in the study of scientific claim verification, and that can inform future work on this task.

Unequal allocation of evidence Fig. 3 shows the distribution of evidence amongst claims in SCIFACT-OPEN. We find that evidence is distributed unequally; half of all ECAPs are allocated to 34 highly-studied claims (12% of all claims in the dataset). We investigated the characteristics of highly-studied claims, and found that they tend to
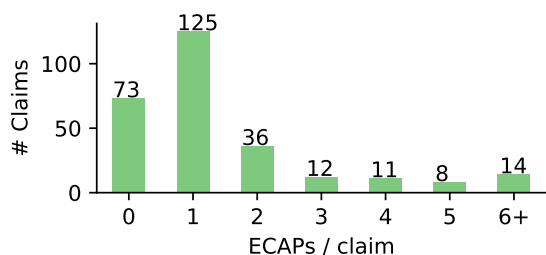
Figure 3: Evidence allocation among claims in SCIFACT-OPEN. The x-axis indicates the number of ECAPs (evidentiary claim / abstract pairs) associated with a given claim, and the y-axis is the number of claims with the corresponding number of ECAPS. For instance, 125 claims are associated with a single evidence-containing abstract.

Claim	ECAPs
Obesity is determined in part by **genetic factors**.	19
Inhibiting **HDAC6** decreases survival of **mice** with **ARID1A mutated tumors**.	0

Table 3: Example of a claim with a number of ECAPs annotated during pooled data collection (top), and another with no new ECAPs (bottom). Well-studied ECAPs tend to be shorter and mention a small number of common entities.

be short and mention a small number of common, well-studied scientific entities. For instance, entities mentioned in well-studied claims ($\geq$ 4 ECAPs) return, on average, 4 times as many documents when entered into a PubMed search, compared to claims with no evidence (detailed results in Appendix B). Table 3 shows an example.

Mismatch in claim and evidence specificity
During evidence collection for SCIFACT-OPEN, annotators reported situations where a claim and abstract exhibited a relationship, but where the claim applied at a different level of *specificity* from the evidence. For instance, in Fig. 1, the claim and refuting evidence discuss the effects of alcohol consumption on *overall* cancer risk, while the supporting evidence indicates that alcohol consumption lowers thyroid cancer risk *in particular*; the supporting evidence is *more specific* than the claim. We also saw cases where the abstract was *more general* than the claim (e.g. claim discusses thyroid cancer, abstract discusses cancer in general), and where the abstract was *closely related* to the claim (e.g. claim discusses thyroid cancer, abstract

Category	ECAPs
Evidence matches claim	115
Evidence more specific than claim	53
Evidence more general than claim	18
Evidence closely related to claim	20

(a) Specificity relationship between claim and evidence, for 206 ECAPs. Specificity mismatches are common, comprising 44% of annotated examples.

Claim: Teaching hospitals provide better care than non-teaching hospitals.
Evidence: Teaching centers . . . prolong survival *in women with any gynecological cancer* compared to community or general hospitals.
Revised claim: Teaching hospitals provide better *gynecological cancer* care than non-teaching hospitals

(b) A CAP where the evidence SUPPORTS a special case of the claim, paired with a *revised* version of the claim that matches the evidence. The claim discusses medical care overall, while the evidence discusses gynecological cancer care specifically.

Table 4: Claim / evidence specificity mismatch in SCIFACT-OPEN.

discusses throat cancer).[5]

Based on this observation, we attempted to quantify the frequency of specificity mismatches. For 206 CAPs in the SCIFACT-OPEN annotation pool, in addition to collecting a SUPPORTS / REFUTES / NEI label, annotators indicated the specificity relationship between claim and abstract, and wrote a *revision* of the claim such that the revised claim matched the specificity of the abstract. These annotations will be released as part of SCIFACT-OPEN.

Table 4a shows counts for different specificity relationships. We find that 91 / 206 (44%) of the examined CAPs exhibit some form of specificity mismatch. Table 4b shows an example where the evidence is more specific than the claim, along with a revised version of the claim that matches the specificity of the evidence. Examples for all specificity relation types — along with analysis showing that mismatches occur in both well-studied and less-studied claims — are included in Appendix B.2. We discuss possible implications of specificity mismatch for future work on scientific claim verification in §7.

Conflicting evidence Conflicting evidence occurs when a single claim is SUPPORTED by at least one ECAP in SCIFACT-OPEN, and REFUTED by another (see Fig. 1). Of the 81 claims in SCIFACT-OPEN with at least 2 ECAPs, 16 of them (20%)

---

[5]In cases of mismatching evidence, we follow the convention used in Thorne et al. (2018) to assign an overall SUPPORTS / REFUTES / NEI label; see Appendix B.2 for details.

Model	SCIFACT-ORIG			SCIFACT-OPEN			
	P	R	F1	P	R	F1	Average Precision
VERT5ERINI	64.0	73.0	68.2	$25.0_{1.9}$	$67.2_{2.9}$	$36.4_{2.2}$	$27.5_{3.2}$
PARAGRAPHJOINT	75.8	63.5	69.1	$54.7_{3.2}$	$46.5_{3.5}$	$50.3_{2.7}$	$40.5_{3.1}$
MULTIVERS	73.8	71.2	72.5	$73.6_{2.9}$	$40.7_{3.3}$	$52.4_{3.0}$	$44.9_{3.7}$
MULTIVERS$_{10}$	63.0	73.0	67.6	$49.6_{3.0}$	$53.0_{3.7}$	$51.3_{2.4}$	$43.4_{3.4}$
ARSJOINT *	72.2	70.3	71.2	$46.1_{2.9}$	$37.6_{3.4}$	$41.4_{2.7}$	-

Table 5: System performance on SCIFACT-OPEN. For comparison, metrics on SCIFACT-ORIG are also reported. Performance is substantially lower on SCIFACT-OPEN relative to SCIFACT-ORIG. Precision, recall, and F1 vary widely by system, based on the negative sampling rate used during training. Subscripts indicate standard deviations over 1,000 bootstrap-resampled versions of the claims in SCIFACT-OPEN (see Appendix C).
*The results for ARSJOINT are not comparable with the other systems, since ARSJOINT was not used for data collection. We did not compute model confidence scores for ARSJOINT; therefore average precision is not reported.

have conflicting evidence. In examining these conflicts, we found that they were often a result of specificity mismatches as shown in Fig. 1 (see Appendix B for additional examples), indicating that modeling evidence specificity represents an important area for future work.

4 Model performance on SCIFACT-OPEN

We evaluate all models from Table 1 on SCIFACT-OPEN. These models represent the state-of-the-art on SCIFACT-ORIG, making them strong baselines to assess the difficulty of our new test collection.

SCIFACT-OPEN is challenging Table 5 shows the performance of all models on SCIFACT-OPEN, as well as on SCIFACT-ORIG for comparison. Due to the wide variation in the precision and recall of different models on SCIFACT-OPEN, we also report average precision, which summarizes performance via the area under the precision / recall curve. We find that models rank similarly on F1 and average precision. Model performance drops by 15 to 30 F1 on SCIFACT-OPEN relative to SCIFACT-ORIG, indicating that all models have trouble generalizing to large corpora unseen during training. PARAGRAPHJOINT, MULTIVERS, and MULTIVERS$_{10}$ all exhibit similar performance (within one standard deviation of each other), while VERT5ERINI performs worse due to low precision.

In Appendix C, we examine model performance on well-studied and less-studied claims separately. We find that higher-recall models tend to perform better on well-studied claims (for which more evidence is available), while higher-precision models perform better on less-studied claims.

Negative sampling affects generalization As mentioned in §2.2, all models except VERT5ERINI
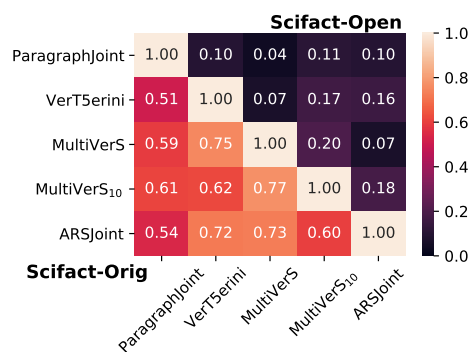

Figure 4: Overlap between the ECAPs predicted by different systems, as measured by Jaccard similarity. Cells below the diagonal show the similarity for abstracts contained in SCIFACT-ORIG, while cells above the diagonal show similarity for abstracts that were added in SCIFACT-OPEN. Overlap is high on abstracts from SCIFACT-ORIG, but much lower when models generalize to documents not seen during training.

were trained with negative sampling. We observe that negative sampling rate has a much larger impact on precision and recall in the open setting than was observed for SCIFACT-ORIG. VERT5ERINI has recall more than double its precision; for MULTIVERS, the situation is reversed. The behavior of MULTIVERS$_{10}$ is much more similar to PARAGRAPHJOINT than MULTIVERS, indicating that negative sampling has a larger impact on model generalization behavior than does model architecture. ARSJOINT is qualitatively similar to PARAGRAPHJOINT and MULTIVERS$_{10}$, but with lower overall performance since its top predictions are not annotated for evidence (see §5.3).

Models have low agreement on SCIFACT-OPEN Fig. 4 shows the overlap among the ECAPs predicted by different systems, measured using Jaccard similarity. Overlap is relatively high (≥ 0.5) for predictions involving abstracts that were found

in SCIFACT-ORIG, and is much lower (≤ 0.2) on abstracts added in SCIFACT-OPEN. From a data collection standpoint, low agreement on SCIFACT-OPEN is a benefit, as it ensures that a diverse set of documents was included in the annotation pool. From a modeling standpoint, it suggests that agreement between existing models when deployed on novel corpora is lower than what has previously been observed. Understanding the differences in the information being identified by each model represents an important direction for future work.

5   Dataset reliability

The total number of annotations collected during pooling (§3.1) is determined by two parameters: the number of annotations per system d, and the number of systems n for which we collect annotations. These parameters must be large enough that increasing them further is unlikely to (1) lead to the discovery of a large number of additional ECAPs or (2) alter the performance metrics of models evaluated on the dataset. Following Zobel (1998), we conduct checks to ensure that conditions (1) and (2) hold for our choices of d and n.

5.1   Annotations per system

To ensure that the number of annotations per system $d = 250$ (also called the *pool depth*[6]) is large enough to ensure reliable evaluation, we examine how much additional evidence is discovered, and how our evaluation metrics change, as d increases from 0 to its final value.

Fig. 5a shows the total number of ECAPS discovered as a function of pool depth. Annotating the 50 most-confident CAPs per system leads to the discovery of 83 ECAPs, while increasing pool depth from 200 to 250 yields 24 new ECAPs—a more than three-fold decrease. This indicates that condition (1) approximately holds; the majority of the evidence in the corpus has been annotated by $d = 250$.

Fig. 5b shows the F1 score of each model as a function of pool depth. While F1 scores change initially, increasing the pool depth from $d = 225$ to $d = 250$ changes the F1 score of each model by less than 2% (see Appendix D for plots). This indicates that condition (2) also holds: further increases to pool depth are unlikely to affect performance metrics. We also find that generalization
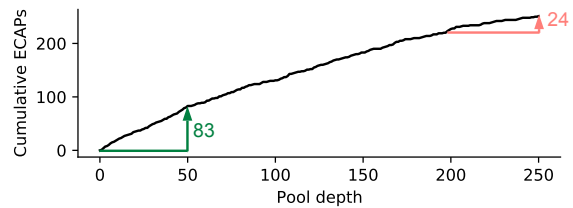
(a) Total number of ECAPs discovered as a function of pool depth. For instance, annotating to a depth $d = 100$ would have resulted in the discovery of roughly 120 ECAPs.
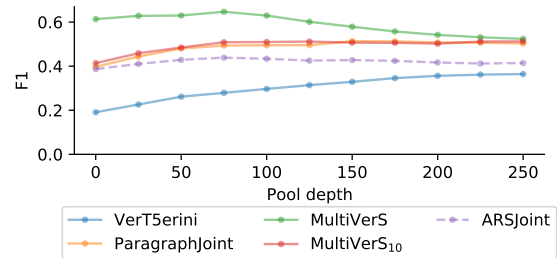
(b) F1 score as a function of pool depth. The blue dot at pool depth 100 indicates that VERT5ERINI would have achieved an F1 score of roughly 30, if annotation had stopped at a depth of 100. Results are ARSJOINT are shown as a dashed line to indicate that this system was not used for data collection.

Figure 5: Effect of pool depth on evidence discovery and evaluation metrics. As pool depth increases, fewer new ECAPs are discovered and F1 score stabilizes.

behavior is influenced more by negative sampling rate than by model architecture. Performance of MULTIVERS decreases with depth, indicating that it was over-fit to the documents in SCIFACT-ORIG, while VERT5ERINI improves with depth. These observations hold if we use average precision rather than F1 to measure performance (Appendix D).
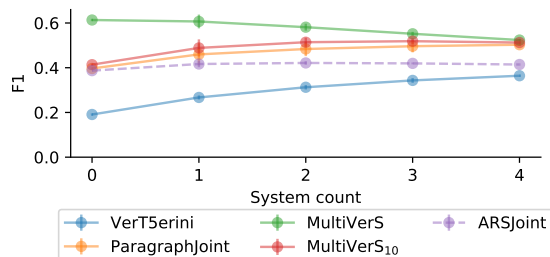
5.2   System count

We repeat the analysis from §5.1, but this time varying the number of systems used for data collection (the *system count*).[7] As was the case for pool depth, Fig. 6a shows that fewer new ECAPs are discovered as more systems' predictions are annotated. Fig. 6b shows that F1 scores stabilize as system count increases, but not as completely as for pool depth; adding a fourth system still leads a 10% change in F1 score for VERT5ERINI and MULTIVERS (Appendix D). Thus, while conditions (1) and (2) are increasingly satisfied as the system count increases, SCIFACT-OPEN would likely benefit from the collection of additional data identified by new models. Unfortunately, unlike pool depth,
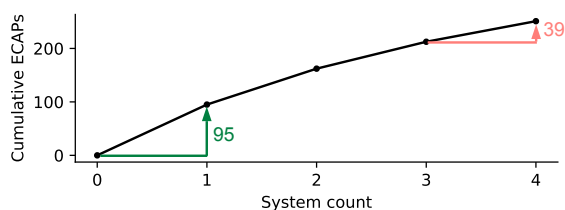
(a) Total ECAPs discovered as a function of system count.

(b) F1 score as a function of system count.

Figure 6: Effect of system count (i.e. number of systems used during pooling) on evidence discovery and evaluation metrics. As in Fig. 5, we see diminishing returns to increasing system count.

the system count that we can achieve is limited by the number of available systems for this task.

5.3 System inclusion

To measure the effect on measured performance of including a given system in the annotation pool, we evaluate each system on the evidence that would have been collected if that system's predictions had not been included. Results are shown in Table 6. All systems except MULTIVERS suffer a roughly 15% drop. When excluded from data collection, PARAGRAPHJOINT and MULTIVERS$_{10}$ both have performance comparable to ARSJOINT. MULTIVERS does not benefit from having its own predictions included, since it was over-fit to SCIFACT-ORIG and struggles to identify new evidence not seen during training. Overall, for fair model comparisons, the performance of new models should be compared against the "Excluded" performance of models used for data collection.

6 Related work

TREC and pooled data collection   Pooling for IR evaluation was popularized by the TREC information retrieval competitions (Voorhees and Harman, 2005), with a number of recent competitions focusing on retrieval in the biomedical domain (Roberts et al., 2020a,b, 2016). Relative to this work, TREC datasets are characterized by a smaller number of queries (or "topics"), a larger number of

Model	Included	Excluded	% Change
VERT5ERINI	36.4	30.5	-16.3
PARAGRAPHJOINT	50.3	42.3	-15.9
MULTIVERS	52.4	51.6	-1.5
MULTIVERS$_{10}$	51.3	43.7	-14.7
ARSJoint	-	41.4	-

Table 6: Change in F1 score when each model is included in the annotation pool, vs. excluded. Omission leads to a performance decrease of roughly 15% for all models except MULTIVERS.

models available for annotation, and a fixed number of annotations per topic (often around 50)— although previous works have proposed strategies to prioritize topics or models for annotation (Zobel, 1998; Cormack et al., 1998). In contrast, to maximize our annotation yield, we collect a variable number of annotations per claim based on model confidence.

Claim verification and revision   Stammbach et al. (2021) studied scientific claim verification against a large research corpus, but simplified the task by evaluating accuracy at predicting a single global truth label per claim, rather than identifying all relevant documents. The Climate-FEVER dataset (Diggelmann et al., 2020) is also open-domain, but assumes a global truth label and verifies claims against Wikipedia, not research papers.

In §3.3, we proposed claim revisions as a solution to claim / evidence specificity mismatch. Claim revision has previously been studied for fact verification over Wikipedia, with the goal of changing the claim from REFUTED to SUPPORTED or vice versa (Thorne and Vlachos, 2021; Schuster et al., 2021; Shah et al., 2020). Previous work has also examined the related task of generating claims based on citation contexts (Wright et al., 2022) and revising questions to match the specificity of answers found in Wikipedia (Min et al., 2020).

7 Discussion & Conclusion

In this work, we introduced a new test collection, SCIFACT-OPEN, to support performance evaluation for open-domain scientific claim verification. The construction of SCIFACT-OPEN was enabled by our adaptation of the pooling strategy from IR for identification and annotation of evidence from a corpus of 500K documents. We hope such methodology can see further usage on other NLP tasks for which exhaustive annotation is infeasible.

In analyzing the evidence in SCIFACT-OPEN

(§3.3), we found that some claims possess a large amount of conflicting evidence, and that evidence may not always match the specificity of the claims as written. We consider two future directions to improve the expressiveness of scientific claim verification. (1) As discussed in §3.3, one could still require systems to label each ECAP, but also to generate a revised claim matching the specificity of each evidence abstract. This output would provide users with fine-grained information indicating the conditions under which an input claim is likely to hold. We release 91 claim revisions, which can be used to facilitate exploratory research in this direction. (2) One could use the evidence identified by a claim verification system as input into a summarization system (DeYoung et al., 2021; Wallace et al., 2021) — potentially using additional quality criteria (e.g. citation count, publication venue) to filter or re-weight the articles included in the summary. This approach has the benefit of providing a concise summary to the user, but there is a greater risk of hallucination (Maynez et al., 2020).

Overall, our analysis indicates that evaluations using SCIFACT-OPEN can provide key insights into modeling challenges associated with scientific claim verification. In particular, the performance of existing models declines substantially when evaluated on SCIFACT-OPEN, suggesting that current claim verification systems are not yet ready for deployment at scale. It is our hope that the dataset and analyses presented in this work will facilitate future modeling improvements, and lead to substantial new understanding of the scientific claim verification task.

8 Limitations

A major challenge in information retrieval is the infeasibility of exhaustive relevance annotation. By introducing an open-domain claim verification task, we are faced with similar challenges around annotation. We adopt TREC-style pooling in our setting with substantially fewer systems than what is typically pooled in TREC competitions, which may lead to greater uncertainty in our test collection. We perform substantial analysis (§5) to better understand the sensitivity of our test collection to annotation depth and system count, and our results suggest that though further improvements are possible, SCIFACT-OPEN is still useful as a test collection and is able to discern substantive performance differences across models. As other models are

developed for claim verification, we may indeed incorporate their predictions in pooling to produce a better test collection.

Through analysis of claim-evidence pairs in SCIFACT-OPEN, we identified the phenomenon of unequal allocation of evidence (§3.3). Some claims are associated with substantially higher numbers of relevant evidence documents; we call these highly-studied claims. In this work, we do not treat these claims any differently than those associated with limited evidence. It could be that highly-studied claims are more representative of the types of claims that users want to verify, in which case we may want to distinguish between these and other types of claims in our dataset, or develop annotation pipelines that would allow us to identify and verify more of these highly-studied claims. In the context of this paper, we derive all claims from the original SCIFACT test collection, and do not provide additional claims.

Finally, we rely on a single retrieval system to identify candidate abstracts. While our analysis indicates that this system identifies the great majority of relevant abstracts (Appendix A.3), future work could extend the dataset collected here by retrieving documents using a wider variety of IR approaches.

Acknowledgments

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *ArXiv*, abs/2004.05150.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An Empirical Investigation of Statistical Significance in NLP. In *EMNLP*.

Daniel Fernando Campos, T. Nguyen, M. Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, L. Deng, and Bhaskar Mitra. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *NeurIPS*.

Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. 1998. Efficient construction of large test collections. In *SIGIR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MS^2: Multi-Document Summarization of Medical Studies. In *EMNLP*.

T. Diggelmann, Jordan L. Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims. In *Tackling Climate Change with ML workshop @ NeurIPS*.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In *ACL*.

V. Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *EMNLP*.

Neema Kotonya and F. Toni. 2020. Explainable Automated Fact-Checking for Public Health Claims. In *EMNLP*.

Xiangci Li, G. Burns, and Nanyun Peng. 2021. A Paragraph-level Multi-task Learning Model for Scientific Fact-Verification. In *Workshop on Scientific Document Understanding @ AAAI*.

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. *ArXiv*, abs/1907.11692.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Michael Kinney, and Daniel S. Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *ACL*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *ACL*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering Ambiguous Open-domain Questions. In *EMNLP*.

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 @EACL*.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, W. Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*.

Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen M. Voorhees, Lucy Lu Wang, and William R. Hersh. 2020a. TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19. *Journal of the American Medical Informatics Association*.

Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, and William R. Hersh. 2016. Overview of the TREC 2016 clinical decision support track. *TREC*.

Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, William R. Hersh, Steven Bedrick, Alexander J. Lazar, and Shubham Pant. 2020b. Overview of the TREC 2020 Precision Medicine Track. *TREC*.

S. Robertson and H. Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic. In *ACL*.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based Fact-Checking of Health-related Claims. In *EMNLP Findings*.

Tal Schuster, Adam Fisch, and R. Barzilay. 2021. Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. In *NAACL*.

Darsh J. Shah, Tal Schuster, and R. Barzilay. 2020. Automatic Fact-guided Sentence Modification. In *AAAI*.

Karen Sparck Jones and C. J. van Rijsbergen. 1975. Report on the need for and provision of an "ideal" information retrieval test collection. In *British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge*.

Dominik Stammbach, Boyang Zhang, and Elliott Ash. 2021. The Choice of Knowledge Base in Automated Claim Checking. *ArXiv*, abs/2111.07795.

Mujeen Sung, Minbyul Jeong, Yonghwa Choi, Donghyeon Kim, Jinhyuk Lee, and Jaewoo Kang. 2022. BERN2: an advanced neural biomedical named entity recognition and normalization tool. *ArXiv*, abs/2201.02080.

Nandan Thakur, N. Reimers, Andreas Ruckl'e, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *NeurIPS*.

James Thorne and Andreas Vlachos. 2021. Evidence-based factual error correction. In *ACL*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for Fact Extraction and VERification. In *NAACL*.

Ellen M. Voorhees and Donna Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval*. MIT press.

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 @ NAACL*.

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 *NAACL Findings*.

David Wadden, Kyle Lo, Lucy Lu Wang, Shanchuan Lin, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *EMNLP*.

Byron C. Wallace, Sayantani Saha, Frank Soboczenski, and Iain James Marshall. 2021. Generating (Factual?) Narrative Summaries of RCTs: Experiments with Neural Multi-Document Summarization. In *AMIA Joint Summits on Translational Science proceedings*.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating Scientific Claims for Zero-Shot Scientific Fact Checking. In *ACL*.

Zhiwei Zhang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. 2021. Abstract, Rationale, Stance: A Joint Model for Scientific Claim Verification. In *EMNLP*.

Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments? In *SIGIR*.

A Dataset construction

A.1 Corpus

We use `2019-09-28` release of the the S2ORC corpus (Lo et al., 2020) as our source of abstracts. We filter for documents whose `mag_field_of_study` field includes at least one of `Medicine` or `Biology`, and which have at least one inbound and one outbound citation; the latter check serves as a basic quality filter to make sure that the article is related to other articles found in the literature. This filtering leaves us with roughly 6.5 million documents, from which we randomly sample our SCIFACT-OPEN corpus of 500K documents (making sure to include the 5K documents from SCIFACT-ORIG). For more information on S2ORC, see `https://github.com/allenai/s2orc`.

A.2 Claims and evidence

Claims As mentioned in §3, we removed 21 claims from the SCIFACT-ORIG test set when constructing SCIFACT-OPEN, leaving 279 claims. Each claim in the dataset is based on a source citation. We removed claims for which we could not find metadata in S2ORC providing information about the source citation — in particular, the article it came from or the year the article was published. No analyses based on this information were included in the final version of this work, but the dataset had already been collected when we realized that this claim filtering step had been unnecessary. Regardless, there is no reason that the availability (or lack thereof) of source metadata would correlate with any linguistic properties of the claims; thus, the omission of these 21 claims should not bias our findings in any way.

Evidence annotation Evidence annotations were performed by three professional annotators with undergraduate degrees in fields related to biology. Before beginning annotations, they participated in a training session with one of the authors to ensure that they understood the task. Annotations are performed as follows. First, every CAP in the annotation pool is assigned randomly to one of the three annotators. The assigned annotator makes a decision on whether the instance clearly does not contain evidence, or whether it might. If it clearly does not, it is marked NEI and annotation stops. If it might contain evidence, the first annotator assigns a label, and a second annotator checks the label to confirm. In the case of disagreement, the two annotators discuss the instance and collectively decide on a final label.

A.3 Retrieval

In §3.1, we chose to retrieve $k = 50$ abstracts per claim; these abstracts were then rank-ordered by model confidence, and the most-confident predictions were annotated for evidence. Using $k = 50$ is justified if worse-ranked retrievals are unlikely to be annotated as ECAPs during the pooling process. Figure 7 confirms that this is indeed the case; the great majority of evidentiary abstracts have retrieval ranks of 20 or better, and very few have ranks worse than 40.
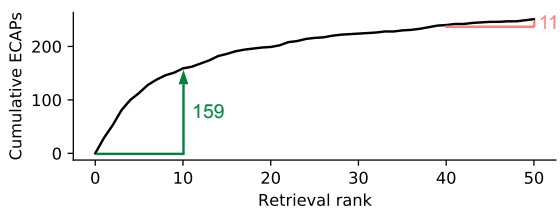
Figure 7: Number of ECAPs discovered as a function of k, the number of abstracts retrieved per claim. The great majority of abstracts judged as ECAPs were ranked among the top 20 retrievals for their respective claims.

A.4 Models

For pooled annotation collection, we used all models achieving state-of-the-art or competitive performance on the SCIFACT leaderboard[8] for which modeling code and checkpoints were available as of early summer 2021, when annotation collection began. The available systems were VERT5ERINI (Pradeep et al., 2021) and PARAGRAPHJOINT (Li et al., 2021) — the two leaders on the SciVer shared task (Wadden and Lo, 2021) — and MULTIVERS (Wadden et al., 2022), formerly called LongChecker. Early in annotation, we noticed that the systems exhibited different precision and recall behavior, and hypothesized that this was due to differences in negative sampling rate. To test this, we also collected annotations with a version of MULTIVERS trained with a negative sampling ratio of 10 (negative sampling ratio is defined in §2.2), referred to as MULTIVERS$_{10}$, and found that this model indeed behaved more like PARAGRAPHJOINT than MULTIVERS in terms of precision and recall. We decided to include MULTIVERS$_{10}$ in the annotation process to increase the diversity of annotation pool. Subsequently, ARSJOINT (Zhang et al., 2021) was released and achieved comparable performance with the four systems used for data collection. We conduct evaluations on this system as well.

System descriptions Given a claim c, all models first retrieve a collection of candidate abstracts a, and then predict labels for each retrieved candidate. In this work, we used the VERT5ERINI retrieval system for all models, since it outperformed the performance of the techniques used with ARSJOINT and PARAGRAPHJOINT. VERT5ERINI first retrieves documents using BM25, then reranks the retrieved documents using a neural reranker trained on MS-MARCO (Campos et al.,

2016). We experimented with using dense retrieval instead (Karpukhin et al., 2020), but found that this did not perform well; similar results were reported in Thakur et al. (2021).

Given a claim c and abstract a, VERT5ERINI selects rationales (evidentiary sentences) from a using a T5-3B model trained on SCIFACT, and then makes label predictions based on the selected rationales using a separate T5-3B model.

PARAGRAPHJOINT and ARSJOINT both encode the claim and full abstract using RoBERTa (Liu et al., 2019), truncating to 512 tokens, and use these representations as the basis for both rationale selection and label prediction. Rationales are predicted based on self-attention over the encodings of the tokens in each sentence, and then a final label is predicted based on self-attention over the representations of the sentences that were selected as rationales.

MULTIVERS encodes the claim and full abstracts in the same fashion as PARAGRAPHJOINT and ARSJOINT, using Longformer (Beltagy et al., 2020) to accommodate long abstracts, and then predicts the label and rationales in a multitask fashion, based on encodings of the leading [SEP] token and sentence separator tokens, respectively.

Negative sampling For scientific claim verification, negative sampling has been performed as follows: for every (c, a) instance in the training data where $y(c, a) \in \{\text{SUPPORTS}, \text{REFUTES}\}$, include r additional $(c, a_i')_{i=1}^r$ instances where $y(c, a_i') = \text{NEI}$ for all i. The irrelevant abstracts a_i' can be sampled randomly from the corpus, or can be chosen to be "hard" negatives; for instance, abstracts a_i' could be chosen which have high lexical overlap with claim c, but which are not annotated as SUPPORTS or REFUTES. Negative sampling has been shown to increase the precision of fact verification models (Li et al., 2021), but comes at the cost of increasing the size of the training dataset (and thus the training time) by a factor of r.

B Additional evidence properties

B.1 Unequal allocation of evidence

Fig. 8a shows the distribution of evidence amongst claims in SCIFACT-OPEN, showing evidence from SCIFACT-ORIG and evidence collected during pooling separately. The majority of claims in SCIFACT-ORIG have one ECAP. Pooling discovered no new ECAPs for the majority of claims in the dataset, and discovered a large amount of
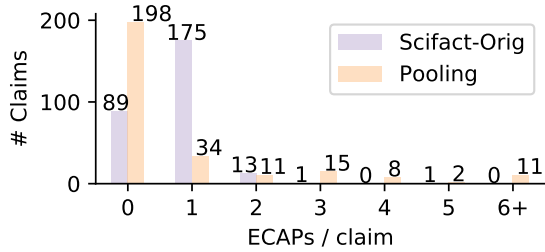
(a) Evidence allocation among claims, showing evidence from SCIFACT-ORIG and evidence collected during pooling separately. The x-axis indicates the number of ECAPs associated with a given claim, and the y-axis is the number of claims with that number of ECAPS.
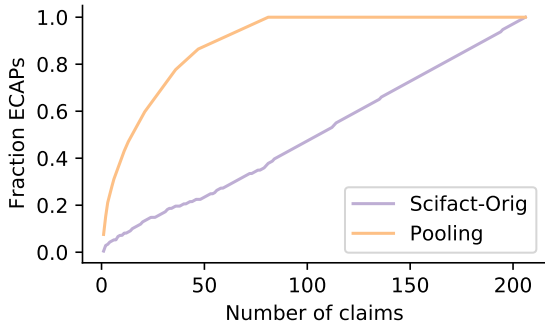
(b) Cumulative distribution of ECAPs across claims.

Figure 8: Distribution of evidence from SCIFACT-ORIG, and from the evidence collected via pooling.

evidence for a small handful of claims. Fig. 8b shows the cumulative distribution of evidence. 14 claims account for 50% of the ECAPs discovered via pooling.

In §3.3, we also observed that well-studied claims tend to be short, and mention a small number of well-studied entities. Table 7 shows these results quantitatively by examining the characteristics of claims for which pooling discovered at least 4 new ECAPs, vs. claims for which it discovered none. Entities mentioned in well-studied claims return, on average, 4 times as many documents when entered into a PubMed search compared with entities mentioned in claims with no new evidence. We use BERN2 (Sung et al., 2022) to identify the entities for this analysis.

B.2 Claim / evidence specificity mismatch

Annotation conventions for mismatched evidence In situations where the evidence in abstract a is more specific or more general than claim c, we follow the convention established in the FEVER dataset (Thorne et al., 2018) to assign a final label $y(c, a)$:

- If a SUPPORTS a special case of c, then assign

	0 ECAPs	≥ 4 ECAPs
Avg. claim length (tokens)	14.1	10.5
Avg. entities / claim	2.3	1.7
Median PubMed results	49K	198K

Table 7: Characteristics of claims for which 0 ECAPs were annotated during pooled data collection, compared to claims with ≥ 4 ECAPs annotated. All differences are significant at the 0.05 level.

Category	ECAPs	
	Well-studied	Less-studied
Evidence matches claim	81	34
Evidence more specific than claim	35	18
Evidence more general than claim	3	15
Evidence closely related to claim	6	14

Table 8: Rates of claim / evidence specificity mismatch for well-studied and less-studied claims.

$y(c, a) = $ SUPPORTS.
- If a SUPPORTS a generalization of c, then assign $y(c, a) = $ NEI.
- If a REFUTES a special case of c, then assign $y(c, a) = $ NEI.
- If a REFUTES a generalization of c, then assign $y(c, a) = $ REFUTES.

Occurrence for well-studied and less-studied claims Table 8 shows rates of claim / evidence specificity mismatch for well-studied and less-studied claims, respectively. Specificity mismatches occur for both types of claims. Interestingly, CAPs where the evidence is more general than the claim occur more frequently for less-studied claims; this likely occurs because less-studied claims are themselves likely to be very specific and cover narrower topics.

Examples In §3.3, we described how the claim and evidence in an ECAP may not have matching levels of specificity. Table 9 provides examples of the different forms of specificity mismatch shown in Table 4.

B.3 Conflicting evidence

Table 10 shows examples of two claims for which conflicting evidence was found in SCIFACT-OPEN.

C Model performance

Uncertainty estimates Table 5 includes uncertainty estimates for performance on SCIFACT-OPEN. We obtain these estimates by computing the

Category	Evidence **matches** claim
Claim	Mitochondria play a major role in calcium homeostasis.
Evidence	Mitochondria . . . are essential organelles responsible for . . . calcium homeostasis.
Revision	
Explanation	No revision necessary; claim and evidence are paraphrases
Category	Evidence **more specific** than claim
Claim	Teaching hospitals provide better care than non-teaching hospitals
Evidence	Teaching centres . . . prolong survival *in women with any gynecological cancer* compared to community or general hospitals.
Revision	Teaching hospitals provide better *gynecological cancer care* than non-teaching hospitals.
Explanation	The evidence refers to *gynecological cancer care* specifically, not care care in general.
Category	Evidence **more general** than claim
Claim	*Somatic missense* mutations in NT5C2 are associated with relapse of *acute lymphoblastic leukemia*.
Evidence	T5C2 mutant proteins show . . . resistance to chemotherapy
Revision	*Mutations* in NT5C2 are associated with relapse of *cancer*.
Explanation	Evidence mentions T5C2 mutations in general, while the claim mentions *somatic missense mutations* specifically. The evidence discusses chemotherapy resistance generally, while the claim discusses relapse of *acute lymphoblastic leukemia* specifically.
Category	Evidence **closely related** to claim
Claim	Near-infrared wavelengths increase penetration depth in *fiberoptic confocal microscopy*
Evidence	Longer wavelength can . . . increase the effective penetration depth of *OCT (optical coherence-domain tomography)* imaging
Revision	Near-infrared wavelengths increase penetration depth in *optical coherence-domain tomography*.
Explanation	The claim discusses *fiberoptic confocal microscopy*. The evidence discusses a different imaging technique, *optical coherence-domain tomography*.

Table 9: Examples of different forms of claim-evidence specificity mismatch. In each example, information specific to claim or evidence is shown in *italics*. The revision re-writes the claim to match the specificity of the evidence.

standard deviation over 1,000 bootstrap-sampled versions of the dataset (Dror et al., 2018; Berg-Kirkpatrick et al., 2012). For a single bootstrap iteration, we resample the claims from the dataset with replacement, and evaluate against the evidence for the sampled claims, weighting the evidence by the number of times each claim was sampled.

Performance for well-studied and less-studied claims   Table 11 shows model performance on well-studied vs. less-studied claims. Higher-recall models tend to perform better on the well-studied claims, since these are the claims where evidence is available in the corpus. Higher-precision models perform better on less-studied claims.

Confusion matrices   Figure 9 shows confusion matrices for all systems. Models rarely confuse SUPPORTS with REFUTES; much more commonly, they either mistake irrelevant abstracts for evidence or fail to identify relevant abstracts.

D   Dataset reliability: Additional experiments

Percentage changes in evaluation metrics   In §5, we examined the effect of pool depth and system count on F1 score. Here, we show the same plots from §5, together with plots showing the *percentage changes* in the F1 score. Results for pool depth are shown in Fig. 10. Results for system count are shown in Fig. 11.

Evaluation using average precision   In §5, we examined the effect of pool depth and system count on F1 score. We perform the same analysis using average precision. Fig. 12 shows the effect of pool depth, and Fig. 13 shows the effect of model count. The qualitative conclusions are the same as for F1. The fact that using F1 and average precision leads to the same conclusions indicates that simply recalibrating each model's classification threshold to adjust for the negative sampling rate used during training would not change the results.

Claim	Bariatric surgery has a deleterious impact on mental health.
SUPPORTS	Our study shows that undergoing bariatric surgery is associated with increases in self-harm, psychiatric service use and occurrence of mental disorders.
REFUTES	Statistical analysis revealed significant improvements in depressive symptoms, physical dimension of quality of life, and self-esteem . . .
Claim	Teaching hospitals provide better care than non-teaching hospitals
SUPPORTS	Teaching centres or regional cancer centres may prolong survival in women with any gynaecological cancer compared to community or general hospitals
REFUTES	Overall [the results] do not suggest that a healthcare facility's teaching status on its own markedly improves or worsens patient outcomes.

Table 10: Examples of two claims with conflicting evidence.

Model	Well-studied				Less-studied			
	P	R	F1	Avg. Precision	P	R	F1	Avg. Precision
VERT5ERINI	31.7	65.2	42.7	30.0	20.9	69.1	32.1	25.5
PARAGRAPHJOINT	57.0	39.6	46.8	40.1	53.2	53.2	53.2	43.8
MULTIVERS	76.1	22.5	34.7	33.6	72.7	58.4	64.8	56.9
MULTIVERS$_{10}$	59.3	42.3	49.4	38.6	44.8	63.5	52.6	52.7
ARSJOINT *	37.5	22.5	28.1	11.3	51.0	52.4	51.7	27.0

Table 11: Performance on SCIFACT-OPEN, for well-studied and less-studied claims. Higher-recall models generally perform better on well-studied claims, while high-precision models perform better on less-studied claims. *The results for ARSJOINT are not comparable with the other systems, since ARSJOINT was not used for data collection.
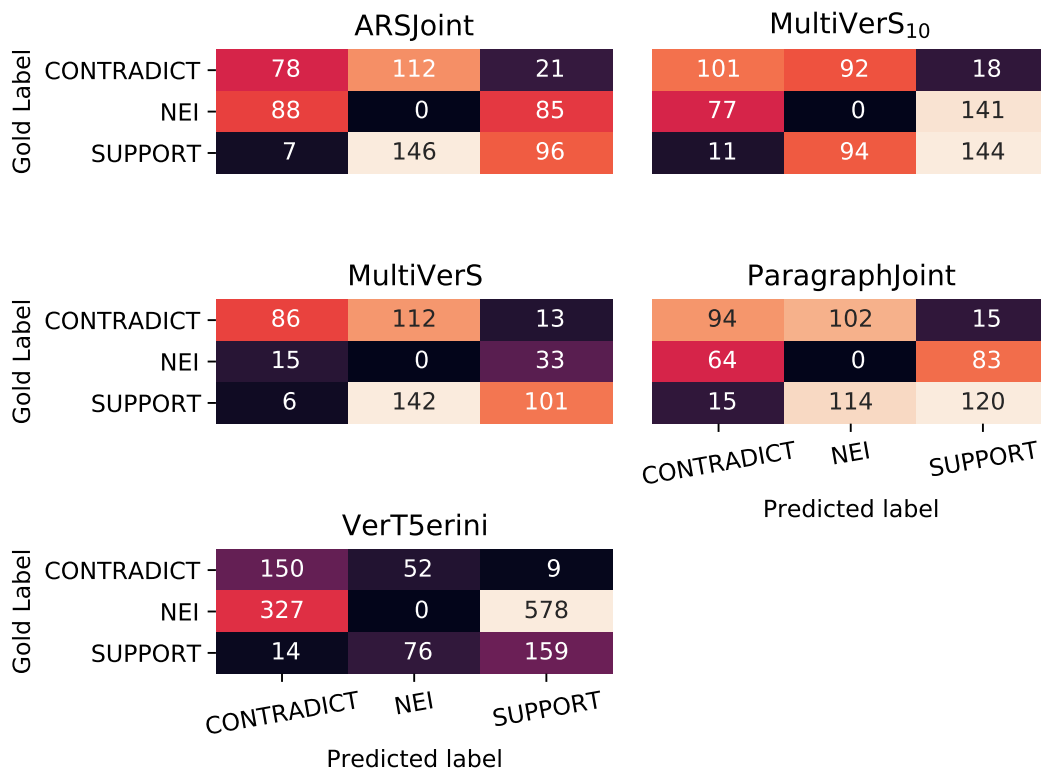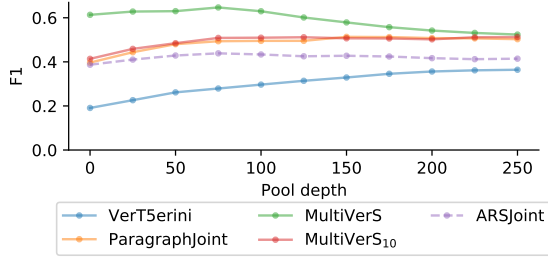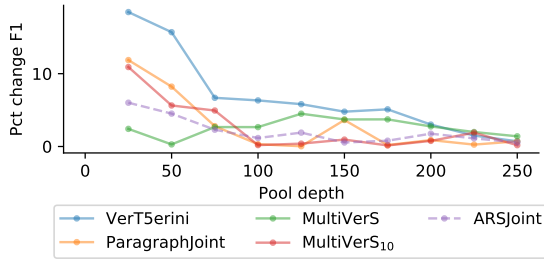

Figure 9: Confusion matrices for model predictions on SCIFACT-OPEN. The {NEI, NEI } cell is 0 because SCIFACT-OPEN is a retrieval task; it's not informative to compute agreement on unlabeled and unpredicted documents in the corpus.
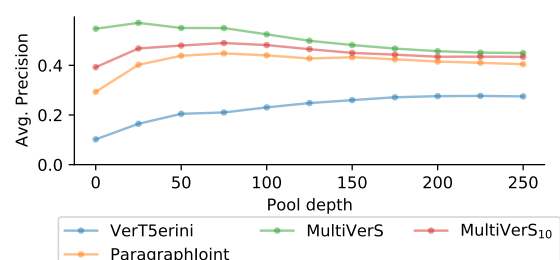
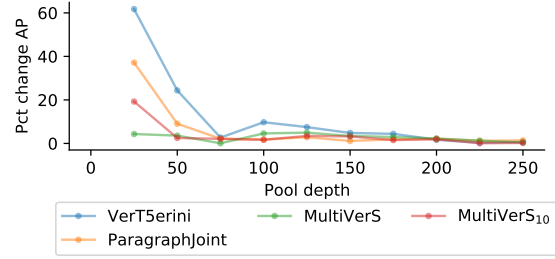(a) F1 score as a function of pool depth. This is the same plot as shown in §5.1

(b) Absolute value of percentage change in F1 score, between adjacent points in the series shown in Fig. 10a. The blue dot at pool depth 100 indicates that the F1 score for VERT5ERINI is increased by roughly 6% when the pool depth increases from 75 to 100. Values close to 0 indicate that collecting additional data does not have an appreciable effect on F1 score.
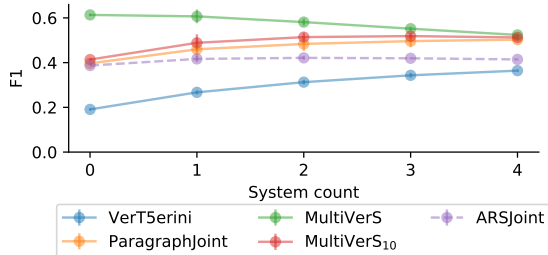
Figure 10: Percentage change in F1 as a function of pool depth.

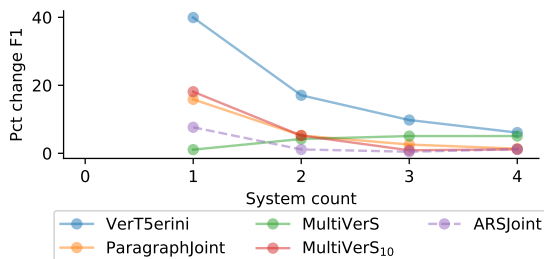(a) Average precision as a function of pool depth.

(b) Absolute value of percentage change in average precision.

Figure 12: Effect of pool depth on model performance, as measured by average precision. We see the same trends as for F1 score. ARSJOINT is not included because computing average precision requires model confidence scores.
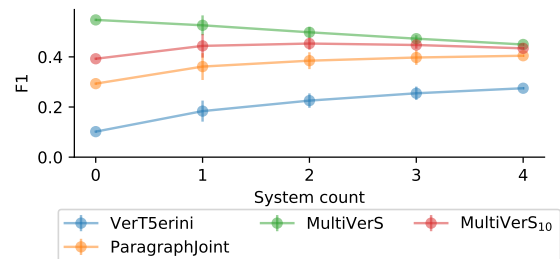

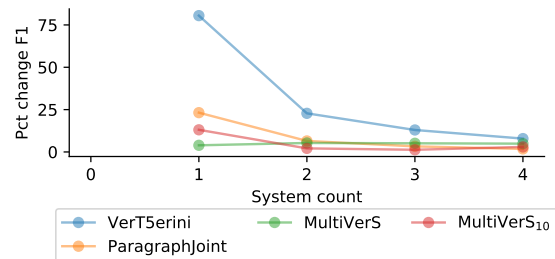(a) F1 score as a function of system count.

(b) Absolute value of percentage change in F1 score, between adjacent points in Fig. 11a.

Figure 11: Percentage change in F1 as a function of system count.

(a) Average precision as a function of system count.

(b) Absolute value of percentage change in average precision.

Figure 13: Effect of system count on average precision. Again, the results here mirror the conclusions drawn using F1 score.