Generating Scientific Claims for Zero-Shot Scientific Fact Checking

Dustin Wright[♮][*]   **David Wadden**[♯]   **Kyle Lo**[♭]

Bailey Kuehl[♭]   **Arman Cohan**[♭]   **Isabelle Augenstein**[♮]   **Lucy Lu Wang**[♭]

[♮]Dept. of Computer Science, University of Copenhagen, Denmark

[♯]University of Washington, Seattle, WA, USA

[♭]Allen Institute for Artificial Intelligence, Seattle, WA, USA

`{dw,augenstein}@di.ku.dk`

`dwadden@cs.washington.edu`

`{kylel,baileyk,armanc,lucyw}@allenai.org`

Abstract

Automated scientific fact checking is difficult due to the complexity of scientific language and a lack of significant amounts of training data, as annotation requires domain expertise. To address this challenge, we propose scientific claim generation, the task of generating one or more atomic and verifiable claims from scientific sentences, and demonstrate its usefulness in zero-shot fact checking for biomedical claims. We propose CLAIMGEN-BART, a new supervised method for generating claims supported by the literature, as well as KBIN, a novel method for generating claim negations. Additionally, we adapt an existing unsupervised entity-centric method of claim generation to biomedical claims, which we call CLAIMGEN-ENTITY. Experiments on zero-shot fact checking demonstrate that both CLAIMGEN-ENTITY and CLAIMGEN-BART, coupled with KBIN, achieve up to 90% performance of fully supervised models trained on manually annotated claims and evidence. A rigorous evaluation study demonstrates significant improvement in generated claim and negation quality over existing baselines.[1]

1 Introduction

Scientific documents contain complex assertions about scientific processes, making it difficult to automate important tasks such as claim extraction and scientific fact checking. Additionally, the collection of manually annotated labels to train models on tasks with scientific data is time consuming and expensive due to the need for domain expertise (Collins et al., 2017; Augenstein and Søgaard, 2017; Lehman et al., 2019; Wadden et al., 2020; DeYoung et al., 2021). As such, methods which require less manual annotation are especially useful in this domain. This work addresses this challenge
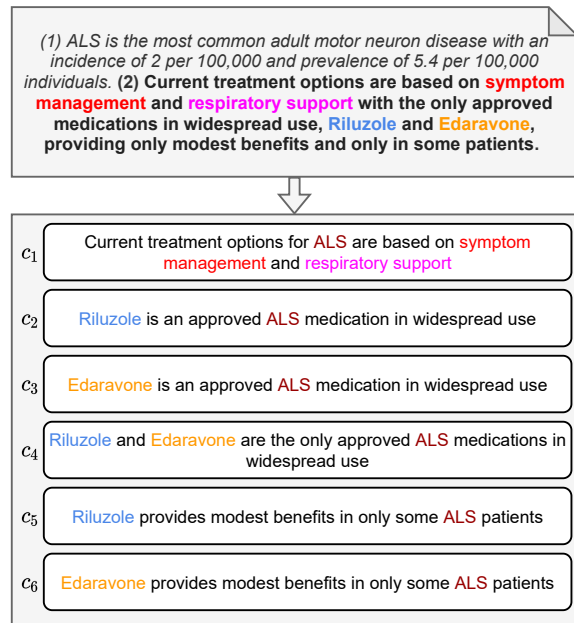

Figure 1: A complex excerpt from Mejzini et al. (2019) (top) and the set of valid claims that can be generated from the bolded sentence (c_1-c_6).

by exploring how automatic generation of scientific claims can assist with dataset creation and zero-shot fact checking in the biomedical domain.

Being able to reduce scientific text to atomic assertions has numerous possible applications, and is known to be helpful for scientific communication and machine processing of scientific concepts (Kuhn et al., 2013). Claim generation can enable zero-shot fact checking, reducing the need for expert-labeled data (Pan et al., 2021), and can be used to expand existing datasets such as Wadden et al. (2020) and Saakyan et al. (2021) without additional manual annotation. In this work we focus on the use of claim generation in scientific fact checking, demonstrating that claim generation enables zero-shot biomedical fact checking.

Generating scientific claims involves distilling a complex scientific sentence into one or more valid claims (see examples in Figure 1). As in previous

---

[*]Work completed while an intern at AI2

[1]Code and data available at: https://github.com/allenai/scientific-claim-generation

work, we focus on biomedical claims as biomedical literature has long been a major focus in scientific natural language processing, as well as scientific fact checking (Saakyan et al., 2021; Wadden et al., 2020; Kotonya and Toni, 2020). While in Wadden et al. (2020), claims were rewritten by domain experts from complex citation sentences (citances), we propose methods for automatically generating claims and claim negations from this source.

Similar to other generation tasks, evaluating the quality of generated output requires multiple judgements beyond the fluency of the generated text, e.g., whether each claim is faithful to the source sentence, and is understandable on its own (Sai et al., 2020). However, there are also other quality attributes that are important to assess specifically for scientific claims, such as whether each claim is atomic or check-worthy (Wright and Augenstein, 2020). Given this, we propose a set of manual evaluation criteria and annotation guidelines for evaluating claim generation (§5.2).

Additionally, when generating claims to build datasets for tasks such as fact checking, a major challenge is creating refuted claims as negative training instances. Previous work has proposed automatic ways of generating refutations based on negating existing claims or creating claim variants via entity-replacement (Pan et al., 2021) and text-infilling using a pre-trained masked language model (Saakyan et al., 2021). We improve upon this by introducing Knowledge Base Informed Negations (KBIN), a principled method to generate refutations that performs entity-replacement using the relations and learned embeddings of entities in a domain-specific knowledge base.

Contributions   In sum, our contributions are:

- The first study on scientific claim generation, comparing both unsupervised (CLAIMGEN-ENTITY) and fully supervised (CLAIMGEN-BART) generation on biomedical text.
- KBIN, a novel method for generating refuted scientific claims which produces more convincing negations than previous work.
- Application of our claim generation methods on zero-shot scientific fact checking resulting in 90% of the performance of a model trained on in-domain manually written claims. Additionally, a rigorous evaluation study showing that CLAIMGEN-BART and KBIN produce significantly higher quality claims and more convincing negations than previous work.

2 Preliminaries

Valid Claims   In this work, we define a *valid claim* as one which is fluent, atomic, de-contextualized, and accurately reflects the meaning of the original sentence. Fluency is concerned with a claim being a generally well-formed English sentence, and atomicity with a claim being a "verifiable statement expressing a finding about one aspect of a scientific entity or process, which can be verified from a single source" (Wadden et al., 2020). De-contextualilzation is concerned with a sentence being interpretable on its own, requiring none of the original surrounding text to resolve aspects of the sentence such as pronouns, abbreviations, etc., and can be handled by either directly de-contextualizing a sentence (Choi et al., 2021) or by ensuring that all of the context sentences are available to a model (Wadden et al., 2021). Check-worthy claims in the wild may not be fluent, atomic, or de-contextualized, however it is useful to generate such claims as they have been shown to be useful for automated processing of science concepts (Kuhn et al., 2013) and scientific fact checking (Wadden et al., 2020).

Scientific Claim Generation   At a high level, scientific claim generation is the task of distilling one or more *valid claims* from one or more sentences concerned with a scientific fact. More specifically, the task is defined as: given a scientific sentence s and optionally additional context sentences X, generate one or more claims $c_i \in C$ which are valid and entailed by s and X. In the context of fact checking, we must generate claims which are either *supported* or *refuted* by the literature, as well as those for which *not enough information* is present to make a veracity judgement, in order that they may be paired with appropriate evidence documents to serve as training data for fact checking systems. As such, we require methods which can take the claims in C which are entailed by the source sentence and generate negations to acquire *refuted* claims.

3 Generating Supported Claims

We experiment with two generation methods designed to produce claims which are *supported* by the source sentence. The first method is an entity-centric unsupervised method adapted from Pan et al. (2021) which requires no <sentence, claim> pairs (CLAIMGEN-ENTITY). We also introduce a new
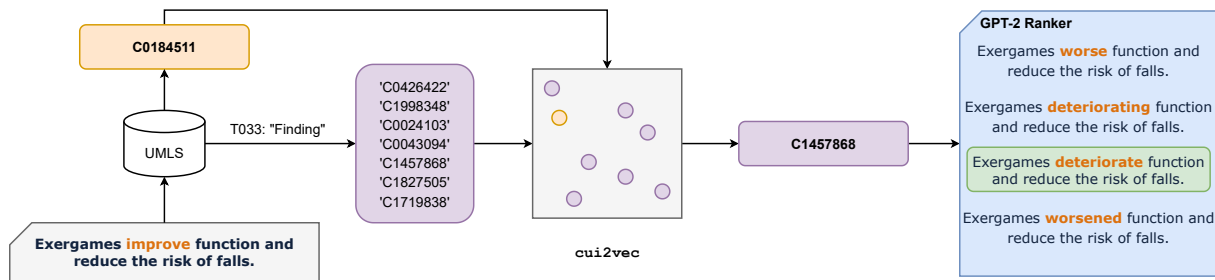
Figure 2: KBIN method. We start with NER and linking to UMLS using scispaCy. We then find the most similar concepts with the same type using `cui2vec`, replace the entity in the source sentence using the canonical name and aliases of similar entities, and rank them using GPT-2. Finally, from the highest ranked replacements, we select the claim which maximizes contradiction with the original claim using an external NLI model.

method that uses BART (Lewis et al., 2020) trained on a small set of <sentence, claim> pairs to directly generate claims (CLAIMGEN-BART). For each sample i, we refer to the input source sentence as s_i, the context sentences as $x_l^{(i)} \in X_i$ and the output claims as C_i consisting of k claims $\{c_1^{(i)} \ldots c_k^{(i)}\}$ Following Wadden et al. (2020), we use citation sentences as unlabelled sentences for generation since these provide a natural link to an evidence document. Various components of our modeling pipelines take advantage of models pretrained on datasets for NER, NLI, QA, and fact-checking. We provide an overview of these datasets in §A.4.

3.1 CLAIMGEN-ENTITY

We adapt the entity-centric method presented in Pan et al. (2021) as an unsupervised claim generation approach. This method has been tested on general domain fact checking, but has not been used for science claim generation and zero-shot scientific fact checking. In particular, we re-implement the base method used for generating supported claims and adapt it to the biomedical domain, substituting in a domain specific model for named-entity recognition. The method consists of the following steps for a given sample i:

1. Run named entity recognition (NER) on the input text to obtain a set of named entities E_i.
2. For each named entity $e_j^{(i)}$, generate a question $q_j^{(i)}$ about that entity which can be answered from s_i.
3. From $q_j^{(i)}$, generate the declarative form of the question to obtain claim $c_j^{(i)}$.

Named Entity Recognition For NER, we employ scispaCy (Neumann et al., 2019), a spaCy[2]

pipeline for scientific NLP. The NER model is trained on the MedMentions dataset (Mohan and Li, 2019), which consists of 4,392 PubMed abstracts exhaustively annotated for mentions of UMLS entities (Bodenreider, 2004).

Question Generation For question generation, we use BART trained on questions from SQuAD (Rajpurkar et al., 2016). As input for training, we encode a concatenation of the context and answer text from a given SQuAD question, and train the model to decode the question. During inference, we concatenate the source sentence s_i and an entity $e_j^{(i)}$ and sample a question $q_j^{(i)}$ for this pair using beam search.

Question to Claim Finally, as in Pan et al. (2021), we use a second BART model to generate declarative claims from questions. We train the model on the QA2D dataset (Demszky et al., 2018), which contains declarative full sentences paired with questions and their answer from SQuAD. The model is trained by encoding a concatenation of the question and answer, and decoding the full declarative sentence. At inference time, we concatenate and encode $q_j^{(i)}$ and $e_j^{(i)}$, and use beam search at the decoder to generate a claim $c_j^{(i)}$.

3.2 CLAIMGEN-BART

We introduce a fully-supervised model for claim generation based on BART trained on <citance, claim> pairs. For this, we use the manual citance re-writes released by the SciFact authors,[3] which consist of citances from scientific papers rewritten as one or more atomic claims which are directly entailed by the citance.

---

[2] https://spacy.io/

[3] https://github.com/allenai/scifact/blob/master/doc/claims-with-citances.md

Algorithm 1 KBIN algorithm

```
 1: function GETNEGATION(c, KB, V, N)
 2:     E ← NER(c)
 3:     C̄ ← []
 4:     for e_j in E do
 5:         u_j ← LINK(e_j)
 6:         R ← KB.siblings(u_j)
 7:         filter(R, KB.type(u_j))
 8:         dist ← cosdist(V[u_j], V[R])
 9:         for r in argsort(dist)[: N] do
10:             A ← KB.aliases(R[r])
11:             T ← replace(c, e_j, a) for a in A
12:             C̄.add(rank_perplexity(T)[0])
13:         end for
14:     end for
15:     return rank_contradiction(c, C̄)[0]
16: end function
```

For training, we encode the citance, as well as the sentences immediately before and after the citance (the context), and train the decoder to generate claims directly. We choose to encode the context as well to help *de-contextualize* generated claims. We concatenate the citance and context using a double pipe (i.e. $X_i | | s_i$), and train the encoder to generate one claim at a time. We use top-k sampling to generate multiple claims, with k set to the number of noun chunks in the original source citance.[4]

4 Knowledge Base Informed Negations

CLAIMGEN-ENTITY and CLAIMGEN-BART only produce claims which are entailed by the source sentence. Additionally, we are interested in producing claim variants which are directly refuted by the original sentence, as these negations are needed when building fact checking datasets and for training fact checking models. Work in Wadden et al. (2020) created these negations manually, and some work has begun to explore automatically generating these negations for scientific claims (Saakyan et al., 2021). To this end, we leverage the availability of large curated biomedical knowledge bases to develop a principled approach to claim variant generation. In particular, we use the UMLS metathesaurus (Bodenreider, 2004), which unifies hundreds of different ontologies in biomedicine, as a source of term replacements for negations.

We provide an overview of the KBIN algorithm

---

[4]We use scispaCy to identify noun chunks

in Algorithm 1 and Figure 2. KBIN works by first performing NER on an input claim c, obtaining entities $\{e_1, \ldots, e_n\} \in E$. For each entity e_j in E, we link the entity to its unique concept u_j in UMLS using the scispaCy entity linker. If the entity is linked, we select all concepts which are siblings to u_j in the concept hierarchy, and which have the same semantic type (e.g. "Clinical Drug"). We rank all selected concepts by their cosine distance to the entity concept using pre-trained UMLS concept vectors, retaining the top 20 closest concepts. For this, we use `cui2vec` (Beam et al., 2020), which contains pre-trained concept vectors for 108,477 concepts from UMLS trained on medical documents from diverse sources.

For each of the related concepts, we generate candidate claim variants by replacing the entity text in the original claim with the canonical name and aliases of the related concept from UMLS. We rank all replacement sentences by their perplexity using a pre-trained GPT-2 model (Radford et al., 2019), keeping the sentence with least perplexity for each replacement. Finally, from among these most fluent sentences, we select the replacement which maximizes the NLI prediction of *contradiction* with the original claim. For this, we use a RoBERTa model (Liu et al., 2019) pre-trained on MNLI (Williams et al., 2018).

5 Experiments

We investigate three primary research questions:

RQ1 Do automatically generated claims enable zero-shot scientific fact checking?

RQ2 What is the percentage of high-quality claims generated using our methods?

RQ3 How does KBIN compare with previous work for claim negation in terms of generating contradictions?

For **RQ1**, we use CLAIMGEN-ENTITY and CLAIMGEN-BART generated claims to train a fact checking model, evaluating on the SciFact dataset (Wadden et al., 2020) and comparing to relevant baselines. To answer **RQ2** and **RQ3**, we design annotation criteria and perform manual evaluations with a group of expert annotators (details in §5.2).

5.1 RQ1: Fact Checking Performance

SciFact Task The SciFact fact verification task consists of: given a claim c and a corpus of scientific abstracts D, retrieve evidence abstracts from

D, predict if the claim is *supported* or *refuted* by those documents or if there is *not enough information (NEI)* to make a prediction, and optionally determine what the rationale sentences are that explain the prediction. Here we focus on the oracle abstract setting of the task, in which gold abstracts are provided to the model and there is no retrieval component. This setup exists in the scientific fact checking literature (Saakyan et al., 2021), and allows us to focus on one component of the fact checking pipeline for evaluating the impacts of claim generation.

Creating Training Data for the Zero-shot Setting We require a set of claim-abstract pairs for training where the abstract either supports, refutes, or does not provide evidence for the given claim. We exploit citation relationships to generate claims paired with potential evidence, using citances from the CiteWorth dataset (Wright and Augenstein, 2021) as source citances for generation. *Supports* claims are produced by directly pairing a generated claim with the abstracts of documents cited by the source citance. For *refutes* claims, we negate a generated claim using KBIN and pair it with the same abstract. For claims labelled *NEI*, we pair the generated claim or negated claim with the abstract of the source document of the citance; the source document is related to the claim but presumably does not directly support or refute the claim given the need for a citation.

Experimental Setup In our experimental setup, we use LongChecker (Wadden et al., 2021), a Longformer (Beltagy et al., 2020) model adapted for scientific fact checking. The model forms its input by concatenating a claim with its evidence abstract, inserting separator tokens between sentences, and uses a classification head to predict the veracity label from the representation of the `[CLS]` token.

We explore several different setups for our training data. As a baseline, we experiment with pre-training only on FEVER claims (Thorne et al., 2018), which are general domain fact checking data based on Wikipedia. We also include an experiment where we manually tune a threshold for the prediction of *NEI* on the SciFact training data, as we saw that the model tends to overpredict this label without any fine-tuning on in-domain data. We also provide an upper bound on performance by fine-tuning on the in-domain train split of SciFact. Finally, we experiment with both

Method	P	R	F1
FEVER only	86.21	11.96	21.01
FEVER + thresh	69.15	66.51	67.80
SciFact (Upper Bound)	77.88	77.51	77.70
CLAIMGEN-ENTITY	**72.86**	69.38	**71.08**
CLAIMGEN-BART	64.09	**79.43**	70.94

Table 1: Results for veracity prediction on the SciFact dataset using different sources of training data.

CLAIMGEN-ENTITY and CLAIMGEN-BART as sources of training data generated from CiteWorth citances, pairing both with KBIN for negations. We note that though CLAIMGEN-BART requires manually re-written claims as training data for generating *supports* claims, it does not use any claims paired with evidence manually labelled for veracity, thus making it zero-shot for the SciFact fact-checking task. In all cases we test on the SciFact dev split. Hyperparameter information, including number of training instances, is given in §A.3, and code and data will be released upon paper acceptance. In all cases, results are reported as macro-F1.

Results Our results on SciFact are given in Table 1. With an upper bound of 77.70 F1, we see that a model fine-tuned on automatically generated claims is able to achieve within 90% of the performance of a model trained on in-domain manually written claims. This is also invariant to the method used to generate claims, as both CLAIMGEN-ENTITY and CLAIMGEN-BART produce similar results. Additionally, both methods provide significant gains over pre-training on FEVER only, especially when no threshold on *NEI* claims is used but also when re-calibrating the model to predict *NEI* less often.

5.2 RQ2: Claim Quality Evaluation

Next, we explore if there are differences between our methods in terms of claim quality and the percentage of valid claims. For this, we ask three expert annotators to manually assess generated claims along a number of quality criteria. One annotator has undergraduate training in the life sciences and graduate training in computer science; the other two annotators have undergraduate training in the life sciences and materials science respectively. We define a set of criteria for evaluation, given in Table 2. These criteria are inspired by the AIDA (Atomic, Independent, Declarative, and Absolute)

Metric	Labels
Fluency	3 - The claim contains no grammatical errors and its meaning can be understood
	2 - The claim contains some grammatical errors but is still understandable
	1- The claim contains many grammatical errors and cannot be understood
De-Contextualized	1 - The claim is interpretable on its own and requires no context; the addition of the original context does not alter the meaning of the claim
	0 - The claim cannot be interpreted in a meaningful way without the original context
Atomicity	1 - The claim is about a single entity/process (atomic)
	0 - The claim is non-atomic and can be broken down into multiple claims
Faithfulness	5 - The claim is correct and fully supported and complete with respect to the original sentence and context
	4 - The claim is correct with respect to the original sentence and context but leaves out information from the original sentence and context
	3 - The claim is related to the original sentence and does not contain incorrect information but is not explicitly stated in the original sentence
	2 - The claim contains explicitly incorrect information relative to the original sentence and context
	1 - The claim has nothing to do with the original sentence

Table 2: Claim quality evaluation metrics and their possible values

framework for scientific claims introduced in Kuhn et al. (2013). They are also based on similar human evaluation criteria used to assess generation quality for related tasks (Sai et al., 2020). We develop an initial set of guidelines for the annotators and conduct two rounds of pilot annotations to improve instructions and increase agreement. For the final evaluation, we generate claims on a set of 100 citances sampled from the CiteWorth dataset (Wright and Augenstein, 2021), which contains citations in context for over 1M citances spanning 10 domains.

We limit the citances to those from papers in biology and medicine to match the domain of Sci-Fact. Annotator agreement is measured as Krippendorff's α (Krippendorff, 2011) on 236 claims for each category except fluency, where we measure the percentage of claims where all annotators agree.[5] The annotators then assess 1,049 total claims (including the 236 shared claims). Each annotator rates all criteria for an individual claim, starting with fluency, then de-contextualized, then atomicity, then faithfulness. We are mainly interested in claim quality and yield, so annotators only annotate "de-contextualized" if the claim is legible (fluency > 1), and only annotate "atomicity" and "faithfulness" if the claim is also de-contextualized (so one is able to discern meaning from the claim). This results in the following rules for acceptable

claims based on the definitions for the labels in each category: Fluency > 1 AND De-Contextualized = 1 AND Atomicity = 1 AND Faithfulness > 3. An acceptable claim is thus legible, meaningful, represents a single aspect of a scientific entity or process, and accurately reflects the information presented in the original citance.

The results of claim quality annotation are given in Table 3. Note that these are on claims generated by CLAIMGEN-ENTITY and CLAIMGEN-BART (see examples in Table 4), and thus are only *supports* claims. We first note that inter-annotator agreement is very high for fluency and moderate across all other criteria. Generated claims are quite fluent across methods, with a small minority of instances being illegible. Unsurprisingly, CLAIMGEN-BART improves over CLAIMGEN-ENTITY across all categories except for atomicity. This intuitively makes sense as CLAIMGEN-ENTITY directly produces claims which are about a single entity. CLAIMGEN-ENTITY yields a higher number of claims per citance as it generates one claim for every entity in the sentence, but the precision of acceptable claims is much lower than that of CLAIMGEN-BART. Thus, there is a trade-off between the two methods between the number of claims generated and their acceptability. While higher yield could lead to higher coverage of claims in the original text, this study is left to future work.

Next, we examine the similarity between generated claims and manually written claims from

---

[5]Fluency agreement is measured in terms of agreement percentage as most ratings are the same (3), thus any disagreements have an oversized influence on α.

Method	Fluency	De-Con. (%)	Atomic (%)	Faithfulness	# Gen	# Accept	P
CLAIMGEN-ENTITY	2.51	55.63	**85.28**	3.54	893	111	12.43
CLAIMGEN-BART	**2.74**	**84.35**	80.65	**4.15**	156	69	**44.23**
α (236 claims)	82.74	64.53	58.71	53.01	-	-	-

Table 3: Average annotation score, agreement, and claim yield for each category. De-contextualized is only annotated if fluency > 1; atomicity and faithfulness are only annotated if fluency > 1 and de-contextualized == 1. # Gen are the total claims generated by the method, and # Accept are the number of acceptable claims generated.

Citance	Generated	Fl,D,A,Fa
Due to its geographic position and geological history, the island of Sardinia is characterized by a remarkable richness of endemic species and represents one of the most prominent biodiversity hotspots in the Mediterranean basin.	The island of Sardinia is characterized by a remarkable richness of endemic species.	3,1,1,5
Frequently reported symptom-eliciting chemicals and environmental agents include fragranted products, motor-vehicle exhaust fumes, cleaning agents, freshly printed papers or magazines, and smoke from wood burners.	Frequently reported symptom-eliciting chemicals and environmental agents are fragranted products.	3,1,1,5
The herbicide inhibits EPSPS (5-enolpyruvylshikimate-3-phosphate synthase) in the shikimate pathway, which has a key role in the biosynthesis of aromatic amino acids and is required for survival of the plant.	The herbicide inhibits EPSPS in the shikimate pathway.	3,1,1,5
Experimental models of OA, such as the intra-articular injection of monosodium acetate (MIA), are associated with joint pathology and pain behaviour comparable to clinical OA.	OA is associated with joint pathology and pain behaviour comparable to clinical OA.	3,1,0,4

Table 4: Sample generated claims with their ratings for (Fl)uency, (D)e-Contextualized, (A)tomicity, (Fa)ithfulness

Method	R-1	R-2	R-L
Entity	47.12	27.63	42.30
BART	**56.58**	**40.12**	**53.38**

Table 5: ROUGE score between generated and manually written reference claims in the SciFact dataset

SciFact. We generate claims for each source citance s_i in the SciFact dev split, and calculate the ROUGE score (Lin, 2004) between each generated claim $c_j^{(i)}$ and each manually written claim $d_k^{(i)}$. From this, we take an average of the max ROUGE score for each generated claim. Formally, given $|C|$ claims we calculate:

$$score = \frac{1}{|C|} \sum_i \sum_j \max_k \text{ROUGE}(c_j^{(i)}, d_k^{(i)})$$

Our evaluation results are given in Table 5. Both methods produce claims which have high overlap with the reference claims, though claims generated directly using BART are significantly closer to the reference claims than those generated using CLAIMGEN-ENTITY. Finally, we note the these scores are in the range of state-of-the-art models

used for paraphrase generation, establishing a solid baseline for this task (Zhou and Bhat, 2021).

5.3 RQ3: Negation Evaluation

Finally, we perform a manual evaluation to compare KBIN against other methods of negation generation. Annotators evaluate negations based on Fluency and Entailment. We adopt the definitions used to annotate the SNLI corpus (Bowman et al., 2015), in which the annotator is given an original claim (premise) and a generated negation (hypothesis) and asked to select from among the following options, including a SKIP option for Fluency:

3 The hypothesis is DEFINITELY FALSE given the premise
2 The hypothesis MIGHT BE TRUE given the premise
1 The hypothesis is DEFINITELY TRUE given the premise
SKIP The hypothesis contains a lot of grammatical errors and cannot be understood

We compare KBIN to two baselines. The first baseline replaces a single entity in the claim with

Original Claim	Method	Generated Negation
Tonic signaling from the SCFV prevents constitutive stimulation.	Entity replace	Tonic signaling from the SCFV under care of respiratory physician (finding) constitutive stimulation.
	Saakyan et al. (2021)	Tonic signaling from the inflammatory stimulation.
	KBIN	Tonic signaling from the SCFV accelerates constitutive stimulation.
Activation of the RAC1 homolog CED-10 kills viable cells in SRGP-1 mutant *Caenorhabditis Elegans*.	Entity replace	Activation of the LASS4 homolog CED-10 kills viable cells in SRGP-1 mutant *Caenorhabditis Elegans*.
	Saakyan et al. (2021)	Activation of the RAC1 homolog CED-10 kills viable cells in SRGP-1 *Helicobacter Elegans*.
	KBIN	Activation of the RAC1 homolog CED-10 mediate viable cells in SRGP-1 mutant *Caenorhabditis Elegans*.

Table 6: Example negations generated using three methods. Span replacements are highlighted in red. In addition to replacing noun phrases, KBIN also has the ability to replace verb phrases as shown in these examples.

Method	Fluency	Entailment		
		3	2	1
Entity replace	83	1	81	1
Saakyan et al. (2021)	83	10	64	9
KBIN	**93**	**15**	75	3

Table 7: Results for manual annotation of claim negations on 100 negations for each method. Fluent claims received annotations other than "SKIP".

a random entity of the same type, similar to the method in Pan et al. (2021). The second is the proposed negation generation method in Saakyan et al. (2021). The method is based on extracting keywords using YAKE (Campos et al., 2020) (an unsupervised method based on statistical text features), replacing those keywords using text infilling with a pre-trained language model, and selecting the replacement with the highest contradiction score using a model pre-trained for NLI. We generate negations for 100 claims using all three methods. For annotation, generated negations from all three methods are aggregated and the order of negation method randomized for each of the 100 claims.

Example negations generated by all three methods are given in Table 6 and annotation results for fluency and entailment are given in Table 7. First, KBIN produces more fluent claims than both baselines. Additionally, KBIN produces more convincing negations on average than both baselines. We observe that the most common operation performed by all three methods is to replace a noun phrase. KBIN has the benefit of being able to replace many entity types corresponding to concepts found in UMLS, which also include verb phrases that encode relations. Finally, KBIN improves over the baseline from Saakyan et al. (2021) by producing fewer claims which are directly entailed by the source claim, i.e., that maintain the original meaning and do not negate the original claim.

5.4 Further Analysis

To give further insight into the quality of claims generated using our methods, we perform an experiment where we train and test models for scientific fact checking using claims only. This "claim-only" experiment helps us assess whether the negation process introduces data artifacts that can be leveraged by the model to predict veracity. We present results from training on claims generated using CLAIMGEN-BART and KBIN, compared against training on the original SciFact training data (which has manually written negations), along with random and majority baselines, in Figure 3.

We observe that there are likely some dataset artifacts in the original SciFact claims that lead to model performance well above the majority and random baselines.[6] This phenomenon has been

---

[6] It is difficult to fully separate the contributions of data artifacts and model performance in this setting, i.e., there is no situation which guarantees *no* undesirable data artifacts. Performance ought to be better than a random baseline in this theoretical setting, due to the pretrained language model likely having had some exposure to the content of the claims during pretraining.
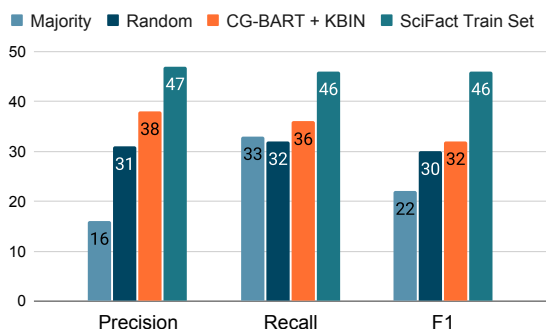
Figure 3: Fact checking performance of models trained only on claims (i.e. no evidence). Training on our generated claims result in performance closer to random (indicating fewer data artifacts) than training on the original SciFact claims.

observed in general domain natural language inference datasets as well (Poliak et al., 2018). Training on claims generated using our methods results in performance that is much more proximal to random performance on the SciFact dev set, indicating that the label-associated bias in the original training data is not present and a possible domain shift between the original SciFact claims and our generated claims. This can further explain some of the performance gap we observe between zero-shot fact-checking and the upper bound of training on manually labeled training data (Table 1).

6   Related Work

Scientific Fact Checking   Our work follows a line of recent literature on scientific fact checking (Wadden et al., 2020). The goal of this task is to determine the veracity of claims related to scientific topics by retrieving appropriate documents from scientific literature, finding evidentiary sentences from those documents, and determining whether claims are supported, refuted, or there is not enough evidence to make a judgement. The task closely resembles the task of general domain fact-checking (Thorne et al., 2018; Augenstein et al., 2019). Well-performing systems on this task use large language models to perform neural document retrieval (Pradeep et al., 2020) or multi-task learning of rationale prediction and stance prediction (Li et al., 2021; Wadden et al., 2021). Recent work on general domain fact checking has also introduced methods for adversarial generation of claims which are particularly difficult to fact-check (Thorne et al., 2019; Atanasova et al., 2020), and for performing the task without any labeled data (Pan et al., 2021).

Our proposed methods extend zero-shot fact checking to the scientific domain, demonstrating that one can achieve 90% of the inference performance of state-of-the-art systems without domain-specific labeled data.

Generating Training Data   Our work is also related to methods for the automatic generation of training data. Generation of synthetic data has been used for multiple tasks, for example question answering (Duan et al., 2017; Riabi et al., 2021), knowledge-base completion (Safavi et al., 2021), and fact-checking (Pan et al., 2021). Most similar to our setting, the COVID-Fact dataset (Saakyan et al., 2021) contains claims related to COVID-19 crawled from Reddit, and is constructed semi-automatically. Claims which are supported by evidence are extracted from Reddit and verified by human annotators, while negations of these claims are generated automatically via masked language model infilling. KBIN improves upon the negation method proposed in this work by leveraging in-domain structured knowledge via UMLS.

7   Conclusion

In this work, we propose the task of scientific claim generation, presenting CLAIMGEN-BART, CLAIMGEN-ENTITY, and KBIN to perform the task. We demonstrate that generated claims can be used to train a model for zero-shot scientific fact checking and obtain within 90% of the performance of a model trained on human-written claims. Through a rigorous user study we demonstrate that CLAIMGEN-BART produces higher quality claims than CLAIMGEN-ENTITY, and that KBIN produces more fluent and more convincing negations than previous work. Work remains to improve claim generation quality and assess the impacts of generated claims in other domains of science, as well as how generated claims can be used in the evidence retrieval component of fact checking systems. We hope that our methods will be used to facilitate future work by enabling faster creation of training datasets and improving the performance of models on the timely and important task of scientific fact checking.

Ethical Considerations

Automated scientific fact checking has great potential value to the scientific community, as well as for addressing phenomenon such as the propagation of scientific misinformation. Our aim in releasing models for scientific claim generation is to improve the generalizability of science fact checking systems in domains with less training resources. When training our fact checking models with generated or synthetic data, there are questions regarding the veracity of the generated data and whether a model trained on inferred labels could produce trustworthy judgments. We hope that by introducing this task and models, we will enable the community to study such questions, while contributing to data curation in a domain in which such curation would normally require significant manual efforts and cost.

References

Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. 2020. Generating Label Cohesive and Well-Formed Adversarial Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3168–3177, Online. Association for Computational Linguistics.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Isabelle Augenstein and Anders Søgaard. 2017. Multi-Task Learning of Keyphrase Boundary Classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 341–346, Vancouver, Canada. Association for Computational Linguistics.

Andrew L. Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin M. Weber, Nathan P. Palmer, Xu Shi, Tianxi Cai, and Isaac S. Kohane. 2020. Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. In *Pacific Symposium on Biocomputing 2020, Fairmont Orchid, Hawaii, USA, January 3-7, 2020*, pages 295–306.

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

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Inf. Sci.*, 509:257–289.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making Sentences Stand-Alone. *Trans. Assoc. Comput. Linguistics*, 9:447–461.

Ed Collins, Isabelle Augenstein, and Sebastian Riedel. 2017. A Supervised Approach to Extractive Summarisation of Scientific Papers. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 195–205, Vancouver, Canada. Association for Computational Linguistics.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming Question Answering Datasets Into Natural Language Inference Datasets. *CoRR*, abs/1809.02922.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MS2: Multi-Document Summarization of Medical Studies. *CoRR*, abs/2104.06486.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question Generation for Question Answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 866–874. 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.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability.

Tobias Kuhn, Paolo Emilio Barbano, Mate Levente Nagy, and Michael Krauthammer. 2013. Broadening the Scope of Nanopublications. In *The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings*, volume 7882 of *Lecture Notes in Computer Science*, pages 487–501. Springer.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pretraining for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. 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 *Proceedings of the Workshop on Scientific Document Understanding co-located with 35th AAAI Conference on Artificial Intelligence, SDU@AAAI 2021, Virtual Event, February 9, 2021*, volume 2831 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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. *CoRR*, abs/1907.11692.

Rita Mejzini, Loren L Flynn, Ianthe L. Pitout, Sue Fletcher, Steve D. Wilton, and Patrick A. Akkari. 2019. ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now? *Frontiers in Neuroscience*, 13.

Sunil Mohan and Donghui Li. 2019. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. In *1st Conference on Automated Knowledge Base Construction, AKBC 2019, Amherst, MA, USA, May 20-22, 2019*.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 319–327. Association for Computational Linguistics.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot Fact Verification by Claim Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 476–483. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 180–191. Association for Computational Linguistics.

Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2020. Scientific Claim Verification with VERT5ERINI. *CoRR*, abs/2010.11930.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2021. Synthetic Data Augmentation for Zero-Shot Cross-Lingual Question Answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7016–7030. Association for Computational Linguistics.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2116–2129. Association for Computational Linguistics.

Tara Safavi, Jing Zhu, and Danai Koutra. 2021. NegatER: Unsupervised Discovery of Negatives in Commonsense Knowledge Bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5633–5646. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2020. A Survey of Evaluation Metrics Used for NLG Systems. *CoRR*, abs/2008.12009.

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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019. The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 1–6, Hong Kong, China. 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 2020, Online, November 16-20, 2020*, pages 7534–7550. Association for Computational Linguistics.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2021. Longchecker: Improving scientific claim verification by modeling full-abstract context. *CoRR*, abs/2112.01640.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Dustin Wright and Isabelle Augenstein. 2020. Claim Check-Worthiness Detection as Positive Unlabelled Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 476–488, Online. Association for Computational Linguistics.

Dustin Wright and Isabelle Augenstein. 2021. Cite-Worth: Cite-Worthiness Detection for Improved Scientific Document Understanding. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1796–1807. Association for Computational Linguistics.

Jianing Zhou and Suma Bhat. 2021. Paraphrase Generation: A Survey of the State of the Art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5075–5086. Association for Computational Linguistics.

Model	Params
RoBERTa	125M
BART	140M
GPT-2	125M
Longformer-SciFact	438M

Table 8: Model sizes.

A Reproducibility

A.1 Computing Infrastructure

All experiments were run on an Amazon Web Services p3.2xlarge instance using a Tesla V100 GPU with 16GB of RAM.

A.2 Number of Parameters per Model

The sizes of each of the models used in this work are given in Table 8.

A.3 Hyperparameters

A.3.1 Fact Checking

SciFact data Learning rate: 1e-5, 5 epochs, gradient accumulation for 8 batches, 1 sample per training batch, 16-bit precision, 809 total claims.

FEVER threshold We tune the NEI threshold on the training set of SciFact, testing values in the range [1e-5, 2e-5, 3e-5, 4e-5, 5e-5, 1e-4, 2e-4, 3e-4, 4e-4, 5e-4, 1e-3, 2e-3, 3e-3, 4e-3, 5e-5, 0.01, 0.12, 0.2, 0.25, 0.4, 0.5, 0.75, 0.8, 0.8, 0.99, 0.999] and find that 5e-5 produces the best result.

CLAIMGEN-BART Learning rate: 2e-6, 5 epochs, gradient accumulation for 8 batches, 1 sample per training batch, 16-bit precision, 1,561 total training claims.

CLAIMGEN-ENTITY Learning rate: 4e-8, 5 epochs, gradient accumulation for 8 batches, 1 sample per training batch, 16-bit precision, 8,592 total training claims.

A.3.2   CLAIMGEN-BART

Learning rate: 2e-5, 3 epochs, linear warmup for 200 steps followed by linear decay, weight decay of 0.01, batch size of 8.

A.4   Description of Datasets

We use a variety of datasets in this study for different components of models, training, and testing. Here we provide a description of each and in which module the dataset is used.

SciFact   The SciFact dataset and rewritten claims used to train CLAIMGEN-BART can be found at https://github.com/allenai/scifact. The dataset consists of 585 original citances with rewritten claims for each of them. Each citance consists of 1-2 rewritten claims. The SciFact rewritten claims are used to train CLAIMGEN-BART for direct claim generation. Additionally, SciFact contains biomedical claims paired with evidence abstracts and veracity labels in {*supports*, *refutes*, *not enough info*} and is split into train, dev, and test sets. We use the train set for supervised fact checking experiments, and the dev set for testing since the test set does not come with labels.

FEVER   FEVER is a general domain fact checking dataset built from Wikipedia. Like SciFact, the dataset consists of claims with paired evidence documents with labels in {*supports*, *refutes*, *not enough info*}. FEVER is used as pretraining data for our fact checking models for zero-shot transfer to biomedical claims. The dataset can be found here https://fever.ai/resources.html.

MedMentions   The MedMentions dataset is a dataset of 4,392 biomedical papers annotated with mentions of UMLS entities. It is used to train the named entity recognition and normalization models used by ScispaCy, which we used for named entity recognition in CLAIMGEN-ENTITY and for normalization in KBIN. The dataset can be found at https://github.com/chanzuckerberg/MedMentions

UMLS   The UMLS meta-thesaurus is a large biomedical knowledge base which unifies hundreds of different ontologies in biomedicine. UMLS is used as the source knowledge base for normalization and candidate selection for KBIN. Additionally, it is the knowledge base used to train cui2vec, which is used for candidate concept selection in KBIN. UMLS can be found here https://www.nlm.nih.gov/research/umls/index.html.

SQuAD   The SQuAD dataset can be found at: https://rajpurkar.github.io/SQuAD-explorer/. SQuAD is used as training data for the question generation module of CLAIMGEN-ENTITY. SQuAD is a question answering dataset which contains data of the form (q, c, a), where q is the question, c is a context document, and a is an answer to the question which can be found in the context.

QA2D   The QA2D dataset can be found at: https://worksheets.codalab.org/worksheets/0xd4ebc52cebb84130a07cbfe81597aaf0/. QA2D is used in the second part of the zero-shot CLAIMGEN-ENTITY model to generate declarative sentences from questions. It consists of data of the form (s, q, a) where q is a question, a is the answer to the question, and s is the declarative form of the question containing the answer.

MNLI   MNLI is a crowd-sourced collection of 433k sentence pairs annotated for textual entailment. In other words, the data consists of pairs (p, h), where p is the premise and h is the hypothesis, and labels in {*entailment*, *contradiction*, *neutral*} which say if the hypothesis entails, contradicts, or is neutral towards the premise. MNLI is used to train a RoBERTa model for entailment, which is used by KBIN to select the best negation among a set of generated claims for a given source citance. The dataset can be found here https://cims.nyu.edu/ sbowman/multinli/