Multi2Claim: Generating Scientific Claims from Multi-Choice Questions for Scientific Fact-Checking

Neşet Özkan TAN Trung Nguyen Joshua Bensemann Alex Yuxuan Peng
Qiming Bao Yang Chen Mark Gahegan Michael Witbrock
The School of Computer Science
University of Auckland
Auckland, New Zealand
{neset.tan, trung.nguyen, josh.bensemann, alex.peng}@auckland.ac.nz
{qiming.bao, yang.chen, m.gahegan, m.witbrock}@auckland.ac.nz

Abstract

In automated scientific fact-checking, machine learning models are trained to verify scientific claims given evidence. A major bottleneck of this task is the availability of large-scale training datasets on different domains, due to the required domain expertise for data annotation. However, multiple-choice question-answering datasets are readily available across many different domains, thanks to the modern online education and assessment systems. As one of the first steps towards addressing the fact-checking dataset scarcity problem in scientific domains, we propose a pipeline for automatically converting multiple-choice questions into fact-checking data, which we call **Multi2Claim**. By applying the proposed pipeline, we generated two large-scale datasets for scientific-fact-checking: **Med-Fact** and **Gsci-Fact** for the medical and general science domains, respectively. These two datasets are among the first examples of large-scale scientific-fact-checking datasets. We developed baseline models for the verdict prediction task using each dataset. Additionally, we demonstrated that the datasets could be used to improve performance measured by weighted $F1$ on existing fact-checking datasets such as SciFact, HEALTHVER, COVID-Fact, and CLIMATE-FEVER. In some cases, the improvement in performance was up to a 26% increase. The generated datasets are publicly available[1].

1 Introduction

Learning to verify the claims in scientific papers and "science releases" (media announcements of scientific findings) is a difficult task for both artificial intelligence (AI) systems and humans. However, this task is crucial because learning to separate verified facts from speculation or falsehoods has important consequences. Success at this task can help the reader understand scientific topics and promote science. Conversely, failure at this task leads to the spread of misinformation and exaggeration, which can cause distortion in scientific communication and undermine public confidence in science. Unfortunately, several studies have revealed that science releases and scientific articles can contain significant exaggeration and misinformation (Woloshin et al., 2009), (West and Bergstrom, 2021), (Sumner et al., 2014), (Woloshin and Schwartz, 2002). This misleading information can directly impact people's lives, as was the case for media releases concerning the COVID-19 pandemic (Roozenbeek et al., 2020). In 2014, some estimates claimed that 40% of the press releases contained exaggerated advice and 33% contained exaggerated claims in science-related news (Sumner et al., 2014). Considering that this problem has not disappeared over the last eight years and has possibly been exasperated by the continued growth of social media and the number of published scientific papers, there is a need for automated systems that can aid both academics and the public in judging the veracity and credibility of scientific claims. In this context, attempts to automate searching for distortions of findings, exaggerations, and misrepresentations

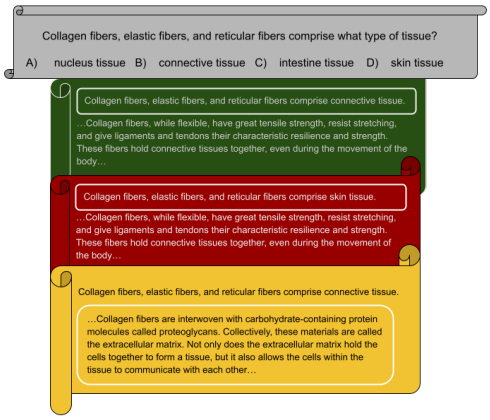

Figure 1: Examples of "supported" (green), "refuted" (red), and "not-enough-info" (yellow) types of claims from the Gsci-Fact dataset which is generated from the original multi-choice question (grey).

[1] https://github.com/taneset/Multi2Claim.

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2652–2664
May 2-6, 2023

may contribute to the trustworthiness verification of science releases and articles.

The fact verification task, commonly known as "fact-checking", is verifying claims in natural language against a collection of information that contains facts. The pipeline for fact-checking usually includes two subtasks: 1) Retrieve-Rank and 2) Veracity Prediction. The Retrieve-Rank operation is typically carried out using information retrieval, and ranking models that rely on a combination of lexical (BM25) and semantic (using word embeddings in pre-trained language models) similarities between the claim and textual evidence candidates (Lin et al., 2020). The veracity prediction task, which has been studied less than the retrieve-rank task, can be seen as a classification task that predicts the type of verdict given a claim-evidence pair. Typically there are three classes of verdicts: "support", "refute", and "not enough information". The veracity prediction problem can also be formulated as a task of identifying textual entailment, in which a model predicts if the provided evidence entails a given claim. Recent work shows that these two subtasks could be combined in an end-to-end manner (Thorne et al., 2021; Wadden et al., 2020a) or could be carried out separately (Saakyan et al., 2021; Diggelmann et al., 2020). Most studies and datasets used in fact-checking are designed to verify claims in general domains such as news, forums, popular Wikipedia passages, and social media posts (Augenstein et al., 2019b; Thorne et al., 2021; Oshikawa et al., 2020; Shahi and Nandini, 2020; Shaar et al., 2020).

It is typically more challenging to automatically verify a scientific claim compared to a claim in the general domain. This is because scientific assertions can be much more complex, and it requires deep domain knowledge to create datasets for scientific claims. The required domain knowledge is a major bottleneck, making the annotation process expensive and time-consuming. As a result of these difficulties, there are only a few scientific-fact-checking datasets in the literature, and the sizes of those datasets are limited. However, the recent deep-learning-based approaches to performing fact-checking require large amounts of annotated training data to generalise well on unseen data. Therefore, there is an urgent need for large-scale scientific-fact-checking datasets, and methods for automatically creating such datasets.

As a step towards addressing the problem of lacking large-scale datasets in scientific-fact-checking, this paper makes the following contributions:

- We constructed a pipeline for generating scientific claims from scientific multiple-choice questions.
- We created two large-scale scientific-fact-checking datasets in the biomedical (150k samples) and general-science domains (32k samples) by applying the proposed pipeline to existing scientific multiple-choice question-answering (QA) datasets.
- We evaluated different pretrained transformer-based models for the verdict prediction task on the generated datasets. The results serve as the initial benchmark on the datasets.
- We showed that the generated datasets can be used to improve the performance on existing scientific-fact-checking datasets.

2 Claim Generation From Multi-Choice Questions

Claim generation can be defined as the process of generating claims that can be classified as "supported", "refuted", or "not enough information" based on evidence in associated texts. A multiple-choice question typically consists of a question, a correct option, and multiple distractors. Some multiple-choice QA datasets even provide an explanation of the correct answer. This section describes a pipeline to automatically generate all three types of claims that are commonly found in the existing fact-checking datasets by taking advantage of such multiple-choice QA datasets. The pseudocode of the generation process is described in Algorithm 1.

2.1 Supported and Refuted Claim Generation

The key to our method of generating supported and refuted claims is a sequence-to-sequence model that can convert question-answer pairs into their declarative forms. For example, the question-answer pair ("Which of the following hormonal activity is expected immediately prior to ovulation?", "LH surge") might be converted into a declarative sentence such as, "LH surge is expected immediately prior to ovulation". In order to achieve this, we adopted the BART (Lewis et al., 2019) model that was trained to convert question-answer pairs in the Stanford Question Answering Dataset

(SQUAD) (Rajpurkar et al., 2016) into their declarative forms, known as QA2D (Demszky et al., 2018). We denote this model as $BART_{QA2D}$.

Our pipeline begins with multiple-choice questions. These questions are typically prepared by domain experts who determine which pieces of knowledge are essential to test learners on within the related field. To create a "supported" claim from a multiple-choice question, we feed the question and the answer (correct option) as input to $BART_{QA2D}$ to generate the declarative sentence. Using this process, we obtain the same number of claims as the number of questions in the multiple-choice QA dataset. The generated claim is then paired with the original explanation of the correct answer, serving as the supporting evidence.

To generate a "refuted" claim type, we make use of the original distractors, carefully selected by domain experts. We assume that any incorrect option (distractor) should be refuted by the original supportive document (explanation), since there is a single correct option for each question. In our LH surge example, the distractors are FSH surge, Progesterone surge, and Estrogen surge. One can generate as many different refuted claims as the number of distractors from a typical multiple-choice question. However, in our implementation, we only generated one refuted claim from a question using the distractor that was the most similar to the correct option. To achieve this, we computed the cosine similarity scores between the embeddings of the correct choice and each of the distractor choices, and then we chose the distractor with the highest score. The embeddings were computed using the Scispacy named-entity-recognition model[2] (Neumann et al., 2019), which was trained on biomedical corpora such as the MedMentions (Murty et al., 2018), and the BioCreative V CDR corpus[3]. This filtering of distractors can help remove distractors that are dissimilar to the correct option and therefore avoid generating claims that can be obviously refuted. The chosen distractor is then fed into the $BART_{QA2D}$ model along with its question to generate a declarative sentence, e.g., "FSH surge is expected immediately prior to ovulation". This generated claim is then paired with the original explanation to the correct answer.

We want to emphasise that generating "refuted" claims is challenging and might require extra ontology-like mechanisms to replace plausible but false notions or entities in order to declare the claim untrue. However, an ontology-like approach requires extensive filtering to ensure the associated document (explanation) is not supporting the replacement and the replacement is meaningfully integrated into the refuted claim. An alternative way of generating a refuted claim is by simply adding "not" to a supported claim. However, exclusively using this method would result in all refuted claims containing "not". This would limit the diversity of the generated claims and might cause the machine learning models to cheat by simply identifying negation.

Algorithm 1 Claim generation from multi-choice question

Require: **Q**uestion (Q), **E**xplanation (E), **A**nswer (A), **D**istractors (D)

function SUPPORTED($q \in Q, e \in E, a \in A$)
▷ q, e and a belong to the same question.
  $c \leftarrow BART_{QA2D}(q, a)$
  return (c, e) ▷ The evidence e supports claim c.
end function
function REFUTED($q \in Q, e \in E, a \in A, D$)
▷ q, e, a and D belong to the same question.
  $\hat{d} \leftarrow \arg\max_{d_i \in D} cosine(a, d_i)$
  $c \leftarrow BART_{QA2D}(q, \hat{d})$
  return (c, e) ▷ Claim c is refuted by e.
end function
function NOT_ENOUGH_INFO($c, e \in E, a \in A, E, A$) ▷ c is generated using SUPPORTED. e and a are its associated explanation and answer.
  $\hat{E} \leftarrow \arg\text{top}10_{e_i \in E, a_i \in A} cosine (SPECTER(a, e), SPECTER(a_i, e_i))$
  ▷ $\hat{E}$ is sorted in descent.
  for $\hat{e}_j$ in $\hat{E}$ **do**
    if $\hat{e}_j \not\supseteq a$ **then**
      return $(c, \hat{e}_j)$ ▷ $\hat{e}_j$ does not contain enough information to make a judgement on c.
    end if
  end for
end function

2.2 Not-Enough-Info Claim Generation

To generate a not-enough-info claim, we replace the explanation of a supported claim with a similar explanation from another claim but without sharing the same key entities or notions (the answer). An

[2]https://allenai.github.io/scispacy
[3]https://www.ncbi.nlm.nih.gov/research/bionlp/Data

example of this is shown in figure 1, where the original supportive document is about collagen fibers and connective tissue, and the replacement is about collagen fibers and extracellular matrix. Crucially, the replacement does not contain information about connective tissue. To help find similar explanations as the one provided, we compute document-level representations of concatenated answer and explanation for all generated claims using SPECTER as introduced in (Cohan et al., 2020). It was shown that SPECTER can create a dense-vector representation for each scientific document in order to capture the relatedness of the documents (Cohan et al., 2020). By using the dense-vector representations of explanations, we retrieve 10 most similar explanations (using cosine similarity) for each claim. Then, we filter out those that contain the key entity of the claim, and we selected the most similar one among them.

2.3 Med-Fact: Medical domain fact-checking dataset

We applied the proposed pipeline to the MedMCQA dataset (Pal et al., 2022) which consists of real-world medical entrance exam questions, answers (which could be multiple or single), and the supporting document for the correct answer. We selected samples whose supporting documents had more than 50 words to ensure a minimal length of the supporting document. Then we only considered multiple-choice questions that had a single correct option. We also dropped the questions, which had the same supportive documents. In the end, we generated 150K claims, including 50K supported, 50K refuted, and 50K not-enough-info claims. Examples of this dataset are given at the end of the Appendix.

2.4 Gsci-Fact: General science domain fact-checking dataset

We used about 13.7K multiple-choice science-exam questions introduced in (Welbl et al., 2017). These questions are about natural sciences such as biology, physics, and chemistry and are created by crowd workers. As was true for Med-Fact, we applied a filtering process to ensure the length of the supporting document and to have a unique supporting document. We generated about 32.2K claims, of which 10.7K are supported, 10.7K are refuted, and 10.7K are not-enough-info. An example is given in Figure 1.

3 Experiments

We conducted experiments to answer the following research questions:

- **Whether the datasets we generated can be used for verdict prediction tasks?**
- **Can the generated datasets improve the models' performance on scientific-fact-checking tasks?**
- **What is the quality of the generated claims?**

We formulate verdict prediction as a multi-class classification task. For a given claim c and a document d, the model must determine a label

$$l(c, d) \in \{\text{supported, refuted, not-enough-info}\}.$$

We concatenate a claim and its document (explanation) together as input, and the model is trained to predict the claim type (supported, refuted, or not-enough-info) in a supervised manner.

3.1 Baselines for Med-Fact and Gsci-Fact

We selected five pre-trained models from the literature as baselines for fact-checking tasks. We fine-tuned the transformer models BERT (Devlin et al., 2018), DeBERTa (He et al., 2020), SciBERT (Beltagy et al., 2019), Longformer (Beltagy et al., 2020), and BioBERT (Lee et al., 2019) for the verdict prediction task. DeBERTa, SciBERT, and BioBERT are descendants of BERT, and DeBERTa has modified attention mechanisms. SciBERT was trained on a large multi-domain corpus of scientific publications, whereas BioBERT was trained on a large-scale biomedical corpus. Longformer is different from the other transformer-based models because it has an efficient attention mechanism that accepts longer input sizes. We used the weighted F_1 metric for evaluation because models can be accurate at predicting a specific label but inaccurate at others, and weighted F_1 can give better insight about performance than accuracy. The weighted-F_1 score is calculated by averaging all per-class F_1 scores while accounting for support for each class, where support refers to the number of actual class occurrences in the dataset. For both Med-Fact and Gsci-Fact datasets, DeBERTa produced the best performance. The complete results for Med-Fact and Gsci-Fact are shown in Table 1.

.

Models	Med-fact	GSci-Fact
BERT	0.70	0.85
DeBERTa	**0.77**	**0.90**
Longformer	0.72	0.86
BioBERT	0.71	0.86
SciBERT	0.65	0.78

Table 1: Weighted F_1 scores of baseline models on verdict prediction task.

3.2 Performance Improvements on Exiting Datasets

We examined whether training models on Med-Fact and Gsci-Fact can improve the performance of verdict prediction on the existing scientific fact-checking datasets. We used two different setups for our experiments. Firstly, we fine-tuned and evaluated models presented in Table 1 on SciFact, HEALTHVER, and CLIMATE-FEVER with three classes (supported, refuted, and not-enough-info). The results are shown in Table 3. We conducted additional experiments to investigate whether our generated fact-checking datasets can improve performance on binary-classification fact-checking datasets such as COVID-Fact. The results for the COVID-Fact dataset are shown in Table 2.

3.2.1 SciFact

According to (Wadden et al., 2020b), the SciFact dataset consists of 1.4K expert-written scientific claims with associated documents (abstracts of the scientific articles) that contain evidence about the claim. The dataset is in the biomedical domain and, is extracted from S2ORC (Lo et al., 2020). The document length is considerably longer than documents in other scientific-fact-checking datasets since it contains the abstracts of scientific papers. Statistics for all the datasets we used in this study are provided in Table 5 in the Appendix.

Although the SciFact dataset is designed for both retrieval (both sentence level and abstract level) and verdict prediction tasks, we only conducted experiments on the verdict prediction task. This set up exists in the literature (Saakyan et al., 2021; Wright et al., 2022). To do that, we used all the associated documents in the datasets for each claim. Since the test set of the dataset is not publicly available, we merged the training and development sets and reserved 10% as a test set for the experiments. We obtained the best weighted F_1 score (0.77) on SciFact using DeBERTa. By fine-tuning models trained on Med-Fact and Gsci-Fact, we improved the weighted F_1 score to 0.86 and 0.83, respectively.

3.2.2 HEALTHVER

HEALTHVER contains health-related claims obtained from sources such as online forums and search engines (Sarrouti et al., 2021). To verify the claims, the top-10 related abstracts were labeled by annotators as "supports", "refutes", "neutral". As with the experiments on SciFact, we used 10% of the dataset as the test set. We obtained results ranging from 0.68 to 0.78 by fine-tuning five pre-trained models on the HEALTHVER dataset. Training first on our Med-Fact and Gsci-Fact datasets consistently increased all models' performance by 0.16 to 0.20 points.

3.2.3 CLIMATE-FEVER

CLIMATE-FEVER consists of claims related to climate change. Like HEALTHVER, they use techniques such as web scraping and using keywords in search engines (Diggelmann et al., 2020). They treated the verdict prediction task as an entailment prediction task to predict one of the labels "SUPPORTS", "REFUTES", or "NOT ENOUGH INFO". During our experiments, we obtained improvement of 0.11 points on the weighted F_1 score using the Med-Fact dataset in the best case. However, The improvements were generally less than that obtained on the previous two datasets. One possible explanation is that the Med-Fact and Gsci-Fact datasets are dominated by biomedical and natural science topics that are not as related to the climate-change domain as the health and medical subject-dominated datasets.

3.2.4 COVID-Fact

We also examined the effect of transferring models trained on the generated dataset to a binary-classification fact-checking dataset such as COVID-Fact. In the COVID-Fact dataset, a claim can either be "supported" or "refuted". These claims have been scrapped from COVID discussions made in online forums such as Reddit. Five pieces of evidence for each claim were collected from Google search results using a cleaning process. We adopted BERT and DeBERTa models in the experiments. We again observed improvement even though the Med-Fact dataset contains many scientific terminologies and jargon, which one cannot expect from COVID-Fact due to its creation process.

Models	BERT	DeBERTa
COVID-Fact	0.60	0.85
Med-Fact + COVID-Fact	0.68^{+8}	0.89^{+4}

Table 2: Comparison of models' performance (Weighted F_1 score) on COVID-Fact.

3.3 Claim Evaluation

We investigated the quality of the generated claims by asking human annotators to evaluate claims from four perspectives: fluency, contextually, faithfulness, and challenge level. We asked six annotators to manually evaluate the claims by following a guideline inspired by (Kuhn et al., 2013; Wright et al., 2022). Table 6 in the Appendix contains information about the guideline, such as definitions of these perspectives and associated scores.

The annotators consisted of Ph.D. students and Ph.D. graduates in the fields of science, medicine, psychology, and computer science. A random sample of 300 examples was taken, with 150 from the Gsci-Fact corpus and 150 from the Med-Fact corpus, ensuring an equal representation of the three claim types: supported, refuted, and not-enough-info. To minimize annotation costs while still leveraging the expertise of each annotator, the sample was divided among eight experts, who were tasked with evaluating the fluency, contextuality, and challenge level of the generated claims. The final scores were obtained by averaging the annotations of all experts on the related dataset.

For the fluency, the annotators found that 98% of the generated claims have no grammatical errors and are clearly understandable for both Gsci-Fact and Med-Fact. The remaining 2% were deemed understandable despite a few grammatical errors and they were equally distributed among three labels for both dataset. The annotators found that 98% of the generated claims in Med-Fact are interpretable without additional context, while 96% are interpretable without additional context in Gsci-Fact. This minor difference could be due to the diverse background of the annotators who evaluated the Gsci-Fact dataset. 75% of the claims in the Gsci-Fact dataset were marked as cannot be answered without the evidence associated with the claims. However, 95% of the claims in the Med-Fact dataset were marked as not verifiable without the associated evidence.

When evaluating the alignment between the assigned labels generated during the generation progress and the faithfulness scores assigned by annotators during the annotation progress, a high degree of agreement was observed. Specifically, annotators concurred with 92% of the "supported" claims in the Gsci-Fact dataset, with corresponding agreement levels of 89% and 88% for the "refuted" and "not-enough-info" claims, respectively. Similar results were obtained for the Med-Fact dataset, with agreement levels of 93%, 91%, and 90% for the "supported", "refuted", and "not-enough-info" claims, respectively. Examples of each type of claim with their evaluation scores can be found in the Appendix, listed in Table 7 for a refuted claim, Table 8 for a supported claim, and Table 9 for a not-enough-info claim.

3.4 Further Analysis

It has been observed that general-domain natural-language-inference datasets can have a significant amount of bias (Poliak et al., 2018), and we wondered if Med-fact and Gsci-Fact contain such a bias. We investigated the claim-only bias because the aim of the fact-checking task is to evaluate the model's ability to examine the semantic relationship between a claim and the supporting data. We tested all the baseline models using just the claim as an input. We discovered that the model's performances (weighted F1 scores) dropped to at most 35%, which indicating that the label-associated bias is not presented. We interpret this result as that the domain-specific evidence associated with the claim is required to make the correct prediction.

4 Related Work

Recent work in automated fact-checking has made progress in the battle against the spread of false information in the news (Pomerleau and Rao, 2017), social media posts, online forums (Vlachos and Riedel, 2014; Mihaylova et al., 2018), and popular Wikipedia articles (Thorne et al., 2018). There has also been substantial work done that uses statements of fact-checking organizations (Augenstein et al., 2019a; Alhindi et al., 2018).

All the work mentioned so far has focused on claims and related documents from the general domain. However, the proposed models and datasets created can be difficult to apply to scientific domains because of the domain mismatch. Limited research has been conducted with datasets and models that focus on verifying claims made in scientific releases against scientific documents. For exam-

Model	SciFact	HEALTHVER	CLIMATE-FEVER
BERT	0.65	0.72	0.58
DeBERTa	0.77	0.78	0.64
Longformer	0.70	0.77	0.63
SciBERT	0.65	0.69	0.52
BioBERT	0.64	0.73	0.51
BERT$_{Med}$	0.78^{+13}	0.91^{+19}	0.62^{+4}
DeBERTa$_{Med}$	0.86^{+9}	0.94^{+16}	0.75^{+11}
Longformer$_{Med}$	0.76^{+6}	0.93^{+16}	0.73^{+10}
SciBERT$_{Med}$	0.68^{+3}	0.88^{+19}	0.55^{+3}
BioBERT$_{Med}$	0.76^{+12}	0.92^{+19}	0.59^{+8}
BERT$_{Gsci}$	0.78^{+13}	0.90^{+18}	0.61^{+3}
DeBERTa$_{Gsci}$	0.83^{+6}	0.92^{+14}	0.70^{+6}
Longformer$_{Gsci}$	0.79^{+9}	0.93^{+16}	0.64^{+1}
SciBERT$_{Gsci}$	0.65^{0}	0.87^{+18}	0.55^{+3}
BioBERT$_{Gsci}$	0.70^{+6}	0.90^{+17}	0.52^{+1}

Table 3: The results of transferring models trained on Med-Fact and Gsci-Fact to SciFact, HEALTHVER and CLIMATE-FEVER datasets. The first five rows show the baseline results (weighted F1) without the transfer. The second set of 5 rows shows the results and improvements from training on Med-Fact first. The last 5 rows show the results of models trained on Gsci-Fact first.

ple, in (Roozenbeek et al., 2020; Saakyan et al., 2021), claims about COVID-19 made on social media platforms were researched. The general health-related claims made in science releases were studied by (Sarrouti et al., 2021). Additionally, claims about climate change and retrieved evidence from Wikipedia were studied by (Diggelmann et al., 2020). However, these models and datasets were designed to verify claims written in less technical language from public science releases and media posts on platforms such as Reddit (Saakyan et al., 2021). We are aware of only one dataset (SciFact) that focuses on claims extracted from scientific articles against scientific documents (Wadden et al., 2020a).

In recent years, a variety of techniques have been developed to enhance fact-checking abilities. One such model is based on multi-layer perceptrons (MLP) (Riedel et al., 2017), and another is based on attention mechanisms (Parikh et al., 2016). Both models were used as baselines in claim verification on the FEVER dataset (Thorne et al., 2018). Additionally, a Graph Neural Network (GNN) based approach (Liu et al., 2020)) was used for propagating nodes represented by evidence (Ye et al., 2020). Semantic role labeling and logical reasoning tools can also improve GNN-based approaches (Chen et al., 2020). In (Zhou et al., 2019), a graph-based evidence aggregating and reasoning (GEAR) framework was employed to aggregate multi-evidence data.

Recently, transformer-based language models have produced the best performance on fact-checking in general and scientific domains. These claim-verification models usually take concatenated claim and evidence pairs and process them with multi-layer transformer-based models to obtain representations for classifying relationships between the claim and the evidence. Pre-trained BERT models have often been used for classification (supported, refuted, and not enough info). For claim verification, BERT-based models are prevalent (Soleimani et al., 2019; Portelli et al., 2020; Chernyavskiy and Ilvovsky, 2019; Nie et al., 2019; Tokala et al., 2019), while Longformer has been used for verdict prediction (Wadden et al., 2020b; Wright et al., 2022).

Advancements in pre-trained transformer-based sequence-to-sequence language models, such as T5 (Raffel et al., 2020) and BART (Lewis et al., 2019)), have allowed researchers to create fully automatic pipelines for claim generation. In particular, negation generation and explanation generation of claims have been studied in the general-domain fact-checking task (Kotonya and Toni, 2020; Thorne et al., 2021). Entity-centric approach (Pan et al., 2021) uses entities in the scientific text to generate claims and has been applied to scientific claim generation by (Wright et al., 2022).

5 Discussion

Because of the annotation cost, most large-scale fact-checking datasets are synthetically generated. For example, the popular fact-checking dataset, Fever (Thorne et al., 2018), contains a limited number of claims annotated by humans. The rest of the claims were synthetically augmented using paraphrasing, negating, and substituting the original claims. However, such methods have been found to weaken results in several studies since real-world claims have different structures than the synthetically augmented claims (Augenstein et al., 2019a; Sarrouti et al., 2021). When generating a refuted claim, simply adding "not" to a supported claim does not always reflect the structure of real-world refuted claims. In our approach, we use the expert-selected distractors to generate refuted claims. The quantity of refuted claims is another weak point of the existing datasets. The amount of "refuted" claims is dramatically less than that of the "supported" claims in the existing datasets. This results in unbalanced fact-checking datasets (see the Table 5 in the Appendix). Both Med-Fact and Gsci-Fact have a balanced class distribution containing the same number of supported and refuted claims.

It is worth mentioning that the supporting documents for the claims that cannot be judged by a given document have been left as empty in (Thorne et al., 2018) and fact-checking datasets that are designed for binary labels ("supported" and "refuted") (Saakyan et al., 2021). In our proposed methods, we were able to retrieve related documents by using their dense-vector representations and consider them as documents for "not-enough-info" type of claims.

Finally, we want to discuss the entity-centric claim generation process (Pan et al., 2021; Wright et al., 2022). The first step of this process is selecting entities in a text to generate claims from them. These entities become the main objects of the generated claims. The weakness of this approach, according to (Pan et al., 2021) is that the generated claims can be superficial, and the verification of these claims can be done without the models' reasoning capabilities or knowledge of common sense. The degree of importance of these entities can differ for a scientific text that contains entities from multiple domains. Since there is no mechanism to indicate how important the selected entity is compared to the other entities in the given text, this approach might result in a poor selection of entities and, therefore, poor claim data. In our proposed method, using the multiple choices/entities of the question likely avoids that problem, because domain experts select the entities in the questions.

6 Conclusion

We have presented Multi2Claim, a pipeline that converts multiple-choice questions to fact-checking datasets. We specifically focus on challenging scientific domains, where the claim verification process can be complicated due to scientific jargon and complex assertions about fields. We presented two large-scale scientific fact-checking datasets created with this pipeline in biomedical (Med-Fact) and general science domains (Gsci-Fact). We testified these dataset for possible biases and the generated datasets were evaluated from various perspectives. Baseline models for these balanced large-scale scientific fact-checking datasets were also presented. We conducted extensive experiments to examine the benefits of the generated datasets on the existing scientific fact-checking dataset, which suffer from low numbers of samples and unbalanced labels. We consistently obtained improvements for all scientific fact-checking datasets, including binary-labeled datasets such as COVID-Fact.

We hope this work will lead to more breakthroughs in scientific fact-checking, which has received little attention due to the expensive and time-consuming annotation process that has to be done by domain experts. We also hope that the proposed pipeline and baseline models will help develop reliable models that will play an essential role in the scientific claim verification process.

Limitations

The proposed method has several limitations. Our claim generation pipeline relies on multiple-choice question-answering datasets since we limited ourselves to reliable and safe generation progress by using human-created scientific questions and answers. Another limitation is the lack of emphasis on retrieving explicit rationals and reasoning over the retrieved rationals due to the high cost of domain-specific rational annotation progress. For future work, we will extend this pipeline to generate claims from plain scientific texts with additional reasoning capabilities.

References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

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

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019b. MultiFC: A Real-World Multi-Domain dataset for Evidence-Based fact checking of claims.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The Long-Document transformer.

Jiangjie Chen, Qiaoben Bao, Changzhi Sun, Xinbo Zhang, Jiaze Chen, Hao Zhou, Yanghua Xiao, and Lei Li. 2020. LOREN: Logic-Regularized reasoning for interpretable fact verification.

Anton Chernyavskiy and Dmitry Ilvovsky. 2019. Extract and aggregate: A novel Domain-Independent approach to factual data verification. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 69–78, Hong Kong, China. Association for Computational Linguistics.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL*.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A dataset for verification of Real-World climate claims.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention.

Neema Kotonya and Francesca Toni. 2020. Explainable automated Fact-Checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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*, pages 487–501. Springer Berlin Heidelberg.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence pre-training for natural language generation, translation, and comprehension.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained transformers for text ranking: BERT and beyond.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Tsvetomila Mihaylova, Preslav Nakov, Lluis Marquez, Alberto Barron-Cedeno, Mitra Mohtarami, Georgi Karadzhov, and James Glass. 2018. Fact checking in community forums.

Shikhar Murty, Patrick Verga, Luke Vilnis, Irena Radovanovic, and Andrew McCallum. 2018. Hierarchical losses and new resources for fine-grained entity typing and linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 97–109, Melbourne, Australia. Association for Computational Linguistics.

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*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. *AAAI*, 33(01):6859–6866.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France. European Language Resources Association.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot fact verification by claim generation.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. 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*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Dean Pomerleau and Delip Rao. 2017. Fake news challenge.

Beatrice Portelli, Jason Zhao, Tal Schuster, Giuseppe Serra, and Enrico Santus. 2020. Distilling the evidence to augment fact verification models. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 47–51, Online. Association for Computational Linguistics.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text.

Benjamin Riedel, Isabelle Augenstein, Georgios P Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the fake news challenge stance detection task.

Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. 2020. Susceptibility to misinformation about covid-19 around the world. *Royal Society open science*, 7(10):201199.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-Fact: Fact extraction and verification of Real-World claims on COVID-19 pandemic.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based Fact-Checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shaden Shaar, Giovanni Da San Martino, Nikolay Babulkov, and Preslav Nakov. 2020. That is a known lie: Detecting previously Fact-Checked claims.

Gautam Kishore Shahi and Durgesh Nandini. 2020. FakeCovid – a multilingual cross-domain fact check news dataset for COVID-19.

Amir Soleimani, Christof Monz, and Marcel Worring. 2019. BERT for evidence retrieval and claim verification.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, Fred Boy, and Christopher D Chambers. 2014. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *BMJ*, 349:g7015.

James Thorne, Max Glockner, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2021. Evidence-based verification for real world information needs.

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

Santosh Tokala, Vishal G, Avirup Saha, and Niloy Ganguly. 2019. AttentiveChecker: A Bi-Directional attention flow mechanism for fact verification. 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 2218–2222, Minneapolis, Minnesota. Association for Computational Linguistics.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. pages 18–22.

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

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020b. Fact or fiction: Verifying scientific claims.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jevin D West and Carl T Bergstrom. 2021. Misinformation in and about science. *Proc. Natl. Acad. Sci. U. S. A.*, 118(15).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Steven Woloshin and Lisa M Schwartz. 2002. Press releases: translating research into news. *JAMA*, 287(21):2856–2858.

Steven Woloshin, Lisa M Schwartz, Samuel L Casella, Abigail T Kennedy, and Robin J Larson. 2009. Press releases by academic medical centers: not so academic? *Ann. Intern. Med.*, 150(9):613–618.

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.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential Reasoning Learning for Language Representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186, Online. Association for Computational Linguistics.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy.

Appendix

A Computing Sources and Experimental Setup

In all of our experiments, we used NVIDIA Quadro RTX 8000 graphics processing unit with 48GB of RAM capacity.

In the verdict prediction experiments, we used the model checkpoints that were provided by Huggingface (Wolf et al., 2019). BERT, SciBERT, and BioBERT have 12 layers and 12 transformer blocks in each layer, the size of the hidden layer is 768. The Longformer model has 12 layer with a hidden dimension of 512. DeBERTa model has a hidden size of 768 and 12 layers.

The exact number of model parameters that we used in this work is shown in the table 4.

Models	Parameters
BERT	109,484,547
Longformer	148,661,763
DeBERTa	184,424,451
SciBERT	109,920,771
BioBERT	108,312,579
BART	139,420,416

Table 4: Parameters per model

B Hyper-parameters

In the generation part, we used maximum 256 for the max length of the sequence to be generated. The number of highest probability vocabulary tokens to keep for top-k-filtering was 10. We returned one of the independently computed returned sequences for each element in the batch. We used the spaCy named entity recognition model from Scispacy to find similarities between the correct option and the other three options[4]. In the not enough info claim type generation, we used the 'allenai-specter'[5] SPECTER model's checkpoint, which was originally stored in the Allen AI repository. For all of the verdict prediction models, we used a variant of the Adam optimizer (AdamW) with a 1e-5 learning rate and other parameters set to default. The epoch number was usually 5, and we did experiments with a range of (3–40) batch numbers with regard to the availability of the GPU. For evaluation, we used the sklearn library's f_1-score function

C Dataset Statistics and Evaluation

[4]https://allenai.github.io/scispacy/

[5]https://huggingface.co/allenai/specter

Datasets	Context	Label distribution	Total size	Query length	Document length
SciFact	Biomedicine	S:556, R:337, N:516	1.4K	12.9	232.9
HEALTHVER	Health	S:4.3K, R:2.8K, N:5.3K	12.5K	18.4	34.8
CLIMATE-FEVER	Climate change	S:654, R:253, N:474, D:154	1.5K	20.5	77.5
COVID-Fact	Covid	S:1.2K, R:2.7K	4K	13.3	77.9
Med-Fact	Biomedical	S:50K, R:50K, N:50K	150K	13.7	125.1
Gsci-Fact	General Science	S:10.7K, R:10.7K, N:10.7	32.2K	12.8	74.1

Table 5: Statistics of the datasets that we considered in this study for scientific-fact-checking task. The letters S, R, N, D stand for supported, refuted, not enough info and disputed claims, respectively. The last two columns show the average lengths of the claims and the documents.

Fluency	3-The claim is free of grammatical errors, and its meaning is clear. 2-The claim is understandable despite some grammatical errors. 1-The claim is incomprehensible. .
Contextuality	1-The claim can be interpreted without any additional context. 0-Without the original context, the claim cannot be interpreted meaningfully.
Faithfulness	1-The claim is correct with respect to the explanation. 2-The claim is incorrect with respect to the explanation. 3-The claim is related to the explanation, but the verdict of the claim cannot be inferred from the explanation. 4-The claim is not related to explanation in any sense.
Challenge	1- I can confidently say whether the claim is correct or incorrect without reading the explanation. 0- I cannot confidently say whether the claim is correct or incorrect without reading the explanation.

Table 6: Manual evaluation criteria for fluency, contextuality, faithfulness and challange.

Claim	Diarrhoea is a common symptom of haloperidol toxicity.
Explanation	Symptoms of haloperidol toxicity are usually due to exaggerated side effects. Most often encountered are: Severe extrapyramidal side effects with muscle rigidity and tremors, akathisia, etc. Hypotension or hypeension Sedation Anticholinergic side effects (dry mouth, constipation, paralytic ileus, difficulties in urinating, decreased perspiration), coma in severe cases, accompanied by respiratory depression and massive hypotension, shock. Rarely, serious ventricular arrhythmia (torsades de pointes), with or without prolonged QT-time Epileptic seizures.
Fluency-3	The claim is free of grammatical errors, and its meaning is clear.
Contextuality-1	The claim can be interpreted without any additional context.
Faithfulness-2	The claim is incorrect with respect to the explanation
Challange-0	I cannot confidently say whether the claim is correct or incorrect without reading the explanation.

Table 7: An example of evaluation of a refuted claim from the Med-Fact dataset.

Claim	Zinc is more easily oxidized than iron.
Explanation	One way to keep iron from corroding is to keep it painted. The layer of paint prevents the water and oxygen necessary for rust formation from coming into contact with the iron. As long as the paint remains intact, the iron is protected from corrosion. Other strategies include alloying the iron with other metals. For example, stainless steel is mostly iron with a bit of chromium. The chromium tends to collect near the surface, where it forms an oxide layer that protects the iron. Zinc-plated or galvanized iron uses a different strategy. Zinc is more easily oxidized than iron because zinc has a lower reduction potential. Since zinc has a lower reduction potential, it is a more active metal. Thus, even if the zinc coating is scratched, the zinc will still oxidize before the iron. This suggests that this approach should work with other active metals. Another important way to protect metal is to make it the cathode in a galvanic cell. This is cathodic protection and can be used for metals other than just iron. For example, the rusting of underground iron storage tanks and pipes can be prevented or greatly reduced by connecting them to a more active metal such as zinc or magnesium. This is also used to protect the metal parts in water heaters. The more active metals (lower reduction potential) are called sacrificial anodes because as they get used up as they corrode (oxidize) at the anode. The metal being protected serves as the cathode, and so does not oxidize (corrode). When the anodes are properly monitored and periodically replaced, the useful lifetime of the iron storage tank can be greatly extended.
Fluency-3	The claim is free of grammatical errors, and its meaning is clear.
Contextuality-1	The claim can be interpreted without any additional context.
Faithfulness-1	The claim is correct with respect to the explanation.
Challange-0	I cannot confidently say whether the claim is correct or incorrect without reading the explanation.

Table 8: An example of evaluation of supported claim from the Gsci-Fact dataset.

Claim	Glycogen phosphorylase requires thiamine pyrophosphate.
Explanation	Glycogen phosphorylase removes glucose as glucose-1-phosphate from glycogen (phosphorolysis). It contains pyridoxal. Formation of branches in glycogen phosphate (PLP) as a prosthetic group. The alpha-1,4 linkages in the glycogen are cleaved and removes glucose units one at a time. Enzyme sequentially hydrolyses alpha-1,4 glycosidic linkages, till it reaches a glucose residue, 3-4 glucose units away from a branch point. It cannot attack the 1,6 linkage at branch point.
Fluency-3	The claim is free of grammatical errors, and its meaning is clear.
Contextuality-1	The claim can be interpreted without any additional context.
Faithfulness-3	The claim is related to the explanation, but the verdict of the claim cannot be inferred from the explanation.
Challange-0	I cannot confidently say whether the claim is correct or incorrect without reading the explanation.

Table 9: An example of evaluation of not-enough-info claim from the Med-Fact dataset.