NewsClaims: A New Benchmark for Claim Detection from News with Background Knowledge

Anonymous ACL submission

Abstract

Claim detection and verification are crucial for news understanding and have emerged as promising technologies for mitigating news misinformation. However, most existing work has focused on *claim sentence* analysis while overlooking crucial background attributes (e.g., claimer, claim objects). In this work, we present NewsClaims, a new benchmark for knowledge-aware claim detection in the news domain. We redefine the claim detection problem to include extraction of additional background attributes related to each claim and release 889 claims annotated over 143 news articles. NewsClaims aims to benchmark claim detection systems in emerging scenarios, comprising unseen topics with little or no training data. To this end, we provide a comprehensive evaluation of zero-shot and prompt-based baselines for NewsClaims.[1]

1 Introduction

The internet era has ushered in an explosion of online content creation, resulting in increased concerns regarding misinformation in associated media forums (e.g., online debates, news articles, social media). A key element of identifying misinformation is detecting the various claims and arguments that have been presented. In this regard, news articles are particularly interesting as they contain claims in various formats, from arguments directly made by journalists to quotations and reported statements made by prominent public figures.

Most current approaches to check-worthiness estimation (Gencheva et al., 2017; Jaradat et al., 2018; Shaar et al., 2021) and argument mining (Eger et al., 2017; Stab et al., 2018) focus on detecting only asserted claims and the premises they are based upon. Beyond identifying which claims to prioritize and the support/refute relations regarding corresponding premises, these approaches largely ignore

News Text: With the coronavirus pandemic continuing to spread around the globe, people are panicked, and they're looking for answers and explanations. *One wild theory that has made its way around the web is that the virus came from space*. Recently, Chandra Wickramasinghe, known for his work in astronomy and astrobiology, spread the idea that the virus was living on a comet and a piece of that space rock may have fallen to Earth.

Topic: Origin of the virus
Stance: Affirm
Claim Object: space
Claimer: Chandra Wickramasinghe

Figure 1: A news article containing a claim regarding the origin of COVID-19 with the claim sentence in italics, claim boundary in red, and claimer in blue. Claimer stance and claim object also indicated.

relevant *background attributes* of the claim (e.g., *claimer*, *object of the claim*) as they have mainly dealt with data written in a collective style (e.g., Wikipedia) or from a monotone viewpoint (Stab and Gurevych, 2014) (e.g., persuasive essays).

However, information domains such as news articles have more complex arguments, requiring a deeper understanding of what each claim is about and where it comes from. These background attributes can be leveraged to build profiles for popular figures in the news, with a list of previously made claims and their most frequent topics. These profiles could be useful for fact-checking organizations to examine how current claims compare to previous ones from the same claimer. Furthermore, all the claims from the same source in a given article can be summarized to present the overall opinion of an entity and increase news comprehension.

To promote research in this direction, we propose expanding the current claim detection task to include extracting more background attributes relating to the claim. Specifically, given a news article, the task is to extract all claims pertaining to a set of topics along with the corresponding claim span, claimer, claimer's stance, and claim object for each claim. In this regard, we release a new evaluation benchmark for claim detection containing

[1]The baselines, the data, the annotation tools, and the evaluation scripts will be made publicly available upon acceptance.

889 claims annotated over 143 news articles. We consider this in an evaluation setting since harmful content[2] can evolve rapidly, requiring systems that are effective under zero/few-shot settings. Figure 1 shows an example from NEWSCLAIMS, including an extracted claim and its background attributes.

In our benchmark, all news articles are regarding the COVID-19 pandemic, motivated by multiple considerations. First, COVID-19 has gained extensive media coverage, with the World Health Organization coining the term *infodemic*[3] to refer to disinformation related to COVID-19 (Naeem and Bhatti, 2020) and suggesting that "fake news spreads faster and more easily than this virus." Second, this is an emerging scenario with limited previous data related to the virus, making it an ideal candidate to evaluate claim detection in a low-resource setting. NEWSCLAIMS is concerning claims about four COVID-19 topics, namely: the origin of the virus, possible cure for the virus, transmission of the virus, and protection from the virus.

NEWSCLAIMS contains primarily an evaluation set with manual annotations, as we aim to study how existing NLP techniques can be leveraged to tackle claim detection in such emerging scenarios with previously unseen topics. We explore multiple zero/few-shot strategies for subtasks including topic classification, stance detection, and claim object detection. This is in line with recent progress in using pre-trained language models in zero-shot (Yin et al., 2019; Liu et al., 2021) and few-shot (Brown et al., 2020; Schick and Schütze, 2021) settings. Such approaches can be adapted to new use cases and problems as they arise without requiring significant additional training data.

Our primary contributions include: (1) we extend the claim detection task to include the background attributes claimer and object of the claim, in addition to claim boundary and claimer stance; (2) we release a manually annotated evaluation benchmark for this new task, NEWSCLAIMS, covering multiple topics related to COVID-19. To the best of our knowledge, this is the first dataset with such extensive annotations for claim detection over news, with 889 claims annotated over 143 news articles; and (3) we provide a comprehensive evaluation of different approaches including multiple zero-shot and prompt-based few-shot learning baselines for various components of our claim detection task.

[2] harmful-content-blog-post
[3] COVID-19 Infodemic

2 Related Work

Argument mining (Palau and Moens, 2009) involves automatically detecting arguments in text. Structured argumentation mining (Stab and Gurevych, 2014; Eger et al., 2017) aims to identify argument components and relations with respect to each other, which differs from the general analysis of opinions in unstructured argumentation mining (Stab et al., 2018). On the other hand, context-dependent claim detection (Levy et al., 2014) involves detecting claims specifically relevant to a predefined topic, whereas Lippi and Torroni (2015) proposed a context-independent claim detection task in which one attempts to detect claims without a specified input topic. Corpus-wide claim detection (Levy et al., 2017) aims to mine arguments from large text corpora to build an argumentative content search engine (Levy et al., 2018). All of these attempts have exclusively focused on identifying arguments about a topic, but do not deal with identifying the background attributes for the claims, such as the claimer and the claim's object.

The claimer detection subtask is related to attribution. Current attribution methods are mainly sentence-level (Pareti, 2016a) or involve only direct quotations (Elson and McKeown, 2010). In contrast, NEWSCLAIMS requires cross-sentence reasoning for identifying the claimer as it may not be present in the claim sentence (see Figure 1).

There has been recent work in addressing misinformation related to claims about COVID-19. Saakyan et al. (2021) proposed a new FEVER-like (Thorne et al., 2018) dataset, where given a claim, the task is to identify relevant evidence and to verify whether the evidence refutes or supports the claim. However, this does not tackle identifying the claims or who is making these claims. There has also been work (Alam et al., 2020; Jiang et al., 2021) on identifying the check-worthiness of tweets related to COVID-19. However, unlike news articles, tweets do not require attribution for claimer idntification.

3 Claim Detection Task

The primary NEWSCLAIMS task is to identify claims related to a set of topics in a news article along with: (1) corresponding background attributes such as the claimer and object of the claim and (2) the claim boundary and stance. In this section, we describe each of the subtasks in detail. Figure 2 shows the expected output from each subtask given a news article as input.

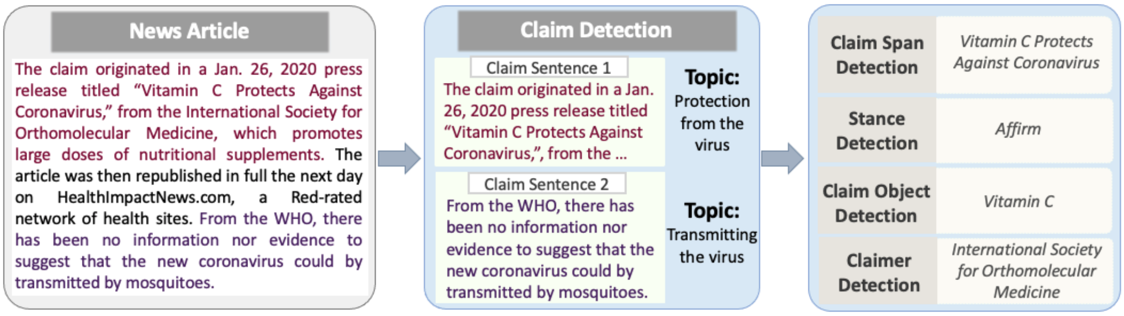


Figure 2: An example demonstrating our proposed claim detection task, and its subtasks. The background attributes such as the claimer, claimer's stance, claim object, and claim span need to be identified for each claim.

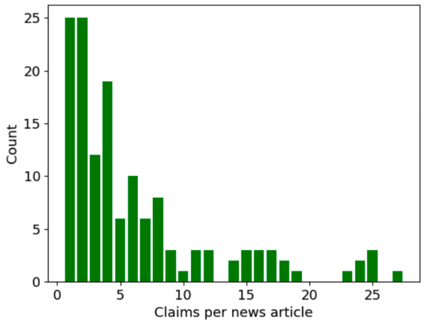
Claim Sentence Detection: Given a news article, the first subtask is to extract claim sentences regarding a set of pre-defined topics. This involves first identifying sentences that contain claims and then selecting those that are related to the topics under consideration. To address misinformation in an emerging real-world setting, we consider the following topics related to COVID-19: **Origin of the virus**: claims related to origin of the virus (i.e., location of first detection, zoonosis, 'lab leak' theories); **Transmission of the virus**: claims related to who/what can transmit the virus or conditions favorable for viral transmission; **Cure for the virus**: claims related to curing the virus, (e.g., via medical intervention after infection); and **Protection from the virus**: claims related to precautions against viral infection.

Claimer Detection: Claims within a news article can come from various types of sources including an entity (e.g., person, organization) or published artifact (e.g., study, report, investigation) when no specific entity is mentioned. In such cases, the claimer identity can usually be extracted from the news article itself. However, if the claim is asserted by the article author or if no attribution is specified or inferable, then the journalist is considered to be the claimer. The claimer detection subtask involves identifying whether the claim is made by a *journalist* or whether it is *reported* in the news article, in which case the source is also extracted. Moreover, sources of such reported claims need not be within the claim sentence. We see that in $\sim 47\%$ of the claims, the claimer span needs to be extracted from outside of the claim sentence. Thus, the claimer detection subtask in our benchmark requires considerable document-level reasoning, thus making it harder than existing attribution tasks (Pareti, 2016b; Newell et al., 2018), which require only sentence-level reasoning.

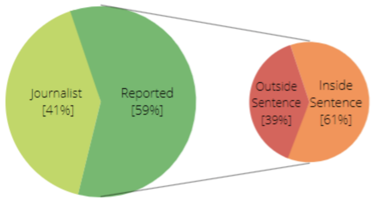
Stance Detection: This subtask involves outputting whether the claimer is asserting (*affirm*) or refuting (*refute*) a claim within the given claim sentence. We note that stance detection in NEWS-CLAIMS differs from the task formulation used in other stance detection datasets (Stab et al., 2018; Hanselowski et al., 2019; Allaway and McKeown, 2020) as it involves identifying the claimer's stance within a claim sentence – whereas prior stance detection tasks, most of which are described in Hardalov et al. (2021), involve identifying the stance for *target–context* pairs. For example, given pairs such as claim–evidence or headline–article, it involves identifying whether the evidence/article at hand supports or refutes a given claim/headline.

Claim Span Detection: Given a claim sentence, this subtask aims to identify exact claim boundaries within the sentence. This claim span covers the actual claim content, usually without any cue words (e.g., asserted, suggested) and frequently a contiguous subspan of the claim sentence. Identifying the precise claim conveyed within the sentence can be useful for latter downstream tasks such as clustering claims and identifying similar or opposing claims.

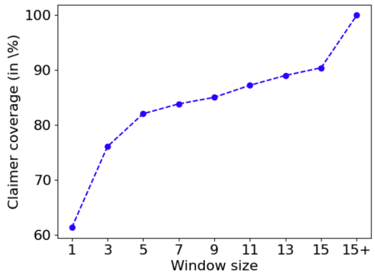
Claim Object Detection: The claim object relates to what is being claimed in the claim sentence with respect to the topic. As an example, for a claim regarding the virus origin, the claim object could be the species of origin in zoonosis claims, or who created the virus in bioengineering claims. Table 1 shows examples of claim objects from each topic. We see that the claim object is usually an extractive span present within the claim sentence. Identifying the claim object helps to better understand claims and can be useful for identifying claim–claim relations, since two claims with the same object are likely to be similar.



(a) Dist. of claim counts per news article.



(b) Dist. of claims from journalists vs those that are reported, along with claimer coverage for reported claims.



(c) Plot showing the coverage of claimer within a window size based on number of sentences around the claim sentence.

Figure 3: Statistics for our claim detection benchmark for (a) counts of claims per news article, (b) claims from journalists vs reported claims, (c) claimer coverage by window size within the news article for reported claims.

Topic	Claim Sentence
Origin	The genetic data is pointing to this virus coming from a **bat reservoir**, he said.
Transmission	The virus lingers in the **air indoors**, infecting those nearby
Cure	**Vitamin C** is an effective treatment for COVID-19.
Protection	Taking a **hot bath** prevents you from getting COVID-19.

Table 1: Examples showing the claim object in bold for claims corresponding to NEWSCLAIMS topics.

4 NEWSCLAIMS Dataset

In this work, we build NEWSCLAIMS, a new benchmark dataset for evaluating the performance of models on different components of our claim detection task. Specifically, we release an evaluation set based on news articles about COVID-19, which can be used to benchmark systems on detecting claim sentences and associated background attributes including claim objects, claim span, claimer, claimer stance. NEWSCLAIMS uses news articles from the LDC corpus *LDC2021E11*, from which we select news articles related to COVID-19. Below, we describe the annotation process (Section 4.1) and provide NEWSCLAIMS statistics (Section 4.2).

4.1 Annotation

Given a news article, we split the annotation process into two phases: (1) identifying claim sentences with their corresponding topics and (2) annotating background attributes for claims identified in the first phase.[4] In the first phase, the interface displays the entire news article with a target sentence highlighted in red. Annotators are asked whether the highlighted sentence contains a claim associated the four pre-defined COVID-19 topics and to indicate the specific topic if that is the case. In the second phase, the interface displays the entire news article with a claim sentence highlighted in red. The annotators are asked to identify the claim boundaries, claim object, and claimer from the news article. The annotators are also asked to indicate the claimer's stance for the claim. We provide a checkbox for the annotators to choose if they think there is no specified claimer, in which case the journalist is considered to be the claimer.

We used Amazon Mechanical Turk (Buhrmester et al., 2011) for annotation, assigning three annotators per example in each phase. In the phase one, only sentences with unanimous support are retained as valid claims. In the second phase, majority voting is used to determine the stance and claimer. Annotators took ∼30 seconds per sentence in the first phase and ∼90 seconds on average annotate the background attributes of a claim in phase two.

4.2 Statistics

NEWSCLAIMS consists of development and test sets with 18 articles containing 103 claims and 125 articles containing 786 claims, respectively. The development set can be used for few-shot labeled examples or for fine-tuning model hyperparameters. Figure 3a shows a histogram of the number of claims in a news article where most news articles contain up to 5 claims and some having more than 10 claims. Claims related to the origin of the virus are the most prevalent, with the respective topic distribution being: 35% for origin, 22% for cure, 23% for protection and 20% for transmission. Figure 3b shows the distribution of claims from journalists vs. reported claims, noting that 41% of the claims are made by journalists with the remaining 59%

[4]Detailed annotation guidelines and screenshots of the interface are provided in Section A.1 in the appendix.

coming from sources mentioned in the news article. Furthermore, for reported claims, the claimer is present outside of the claim sentence 39% of the time, demonstrating the document-level nature of this task. Figure 3c shows the claimer coverage (in %) based on a window around the claim by the number of sentences and indicates that document-level reasoning is required to identify the claimer, with some cases even requiring inference beyond a window size of 15. Note that the 61% inside-sentence coverage in Figure 3b corresponds to a window size of 1 in Figure 3c.

5 Baselines

In this section, we describe various zero-shot and prompt-based few-shot baselines for the claim detection subtasks outlined in Section 3. We describe a diverse set of baselines with each chosen to be relevant in an evaluation-only setting.

5.1 Claim Detection

Given a news article, the task is to detect all the sentences that contain claims relating to a pre-defined set of topics regarding COVID-19. We employ a two-step procedure that first identifies sentences that contain claims and then selects those related to the coronavirus.

Step 1. ClaimBuster: To identify sentences containing claims, we use ClaimBuster (Hassan et al., 2017),[5] a claim-spotting model trained on a dataset of check-worthy claims (Arslan et al., 2020). The model outputs a check-worthiness score for each sentence in the input news article, which we use to select sentences that contain claims. As ClaimBuster has no knowledge about topics, we use zero-shot topic classification, as described below.

Step 2. ClaimBuster+Zero-shot NLI: Following Yin et al. (2019), we use pre-trained NLI models as zero-shot text classifiers: we pose the claim sentence to be classified as the NLI premise and construct a hypothesis from each candidate topic. Figure 4a shows the hypothesis corresponding to each of the topics. We then use the entailment score for each topic as its topic score and choose the highest topic score for threshold-based filtering.

5.2 Claim Object Detection

Given the claim sentence and topic, claim object detection seeks to identify what is being claimed about the topic, as shown in Table 1. We explore this subtask in both zero-shot and few-shot settings by converting it into a prompting task for pre-trained language models as described below:

In-context learning (few-shot): This setting is similar to (Brown et al., 2020), where the few-shot labeled examples are inserted into the context of a pre-trained language model. The example for which a prediction is to be made is included as a prompt at the end of the context . We refer the reader to Section A.3 in the appendix for an example. We use GPT-3 (Brown et al., 2020) as the language model in this setting.

Prompt-based fine-tuning (few-shot): Following Gao et al. (2021), we fine-tune a pre-trained language model (base T5 (Raffel et al., 2020)) to learn from a few labeled examples. Examples are converted into a prompt with a format similar to the language model pre-training, which for this model involves generating the target text that has been replaced with a <MASK> token in the input. Thus, we convert the few-shot data into such prompts and generate the claim object from the <MASK> token. For example, given the claim sentence: *Research conducted on the origin of the virus shows that it came from bats*, and its topic (origin of the virus), the prompt would be: *Research conducted on the origin of the virus shows that it came from bats. The origin of the virus is <MASK>.*

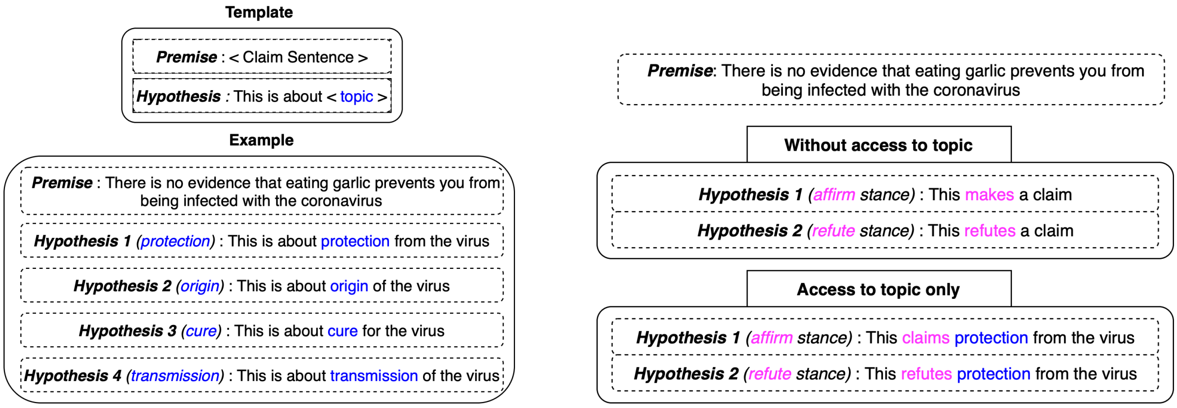
Prompting (zero-shot): We consider the language models that were used in few-shot settings above with the same prompts but in zero-shot settings here. In this case, GPT-3 is not provided with any labeled examples in the context and T5 is used out-of-the-box without any fine-tuning.

5.3 Stance Detection

Given the claim sentence, stance detection identifies if the claimer is asserting or refuting a claim.

Zero-shot NLI: We again leverage NLI models for zero-shot classification. Here, we construct a hypothesis for the *claim* and the *refute* labels and take the stance corresponding to a higher entailment score. We consider two different settings while constructing the hypothesis based on claim topic access. Example hypotheses for each setting are shown in Figure 4b.

[5] https://idir.uta.edu/claimbuster/api/



(a) zero-shot NLI for topic classification

(b) zero-shot NLI for stance detection

Figure 4: Diagram (a) shows the template and an example for leveraging a pre-trained NLI model for zero-shot topic classification; the topic corresponding to the hypothesis with the highest entailment score is taken as the claim sentence topic. Diagram (b) shows examples for leveraging a pre-trained NLI model for zero-shot stance detection. Each example shows how the hypothesis is constructed based on the class label (in pink) and the topic (in blue).

5.4 Claim Span Detection

Given a claim sentence, claim span detection identifies the exact claim boundaries within the sentence.

Debater Boundary Detection: Our first baseline uses the claim boundary detection service from the Project Debater[6] APIs (Bar-Haim et al., 2021). This system is based on BERT-Large, which is further fine-tuned on 52K crowd-annotated examples mined from the Lexis-Nexis corpus.[7]

PolNeAR-Content: Our second baseline leverages PolNeAR (Newell et al., 2018), a popular news attribution corpus of annotated triples comprising the *source*, *cue*, and *content* for statements made in the news. We build a claim span detection model from this by fine-tuning a BERT-large model (Devlin et al., 2019) to identify the *content* span, with a start classifier and an end classifier on top of the encoder outputs, given the sentence as an input.

5.5 Claimer Detection

This subtask identifies if the claim is made by the journalist or a reported source, in addition to identifying the mention of the source in the news article.

PolNeAR-Source: We leverage the PolNeAR corpus to build a claimer extraction baseline. Given a statement, we use the *source* annotation as the claimer and mark the *content* span within the statement using special tokens. We then fine-tune a BERT-large model to extract the source span from the statement using a start classifier and an end classifier over the encoder outputs. During evaluation, we use the news article as an input, marking the claim span with special tokens and using the sum of start and end classifier scores as a claimer span confidence score. This is thresholded to determine if the claim is from the journalist, with the claimer span used as output for reported claims.

SRL: We build a Semantic Role Labeling (SRL) baseline for claimer extraction. SRL outputs the verb predicate-argument structure of a sentence such as who did what to whom. Given the claim sentence as an input, we filter out verb predicates that match a pre-defined set of cues [8] (e.g., *say*, *believe*, *deny*). Then, we use the span corresponding to the ARG-0 (agent) of the predicate as the claimer. AS SRL works at the sentence level, this approach cannot extract claimers outside of the claim sentence. Thus, the system outputs *journalist* as the claimer when none of the verb predicates in the sentence matches the pre-defined set of cues.

6 Experiments

In this section, we empirically evaluate various zero-shot and few-shot approaches on multiple components of our claim detection task, reporting results on the test set. To estimate upper bounds, we also show human performance for each subtask computed over 10 news articles that are randomly

[6]Project Debater
[7]http://www.lexisnexis.com/en-us/home.page

[8]We refer the reader to Section A.2 in the appendix for the complete set of cues.

sampled, with 70 claims in total.

6.1 Claim Detection

Setup: For zero-shot MNLI, we use BART-large[9] (Lewis et al., 2020) trained on the MultiNLI corpus (Williams et al., 2018). ClaimBuster and the topic-filtering thresholds are tuned on the development set. Evaluation is quantified by precision, recall and F1 scores for the filtered set of claims relative to the ground-truth annotations.

Results and Analysis: Table 2 shows the performance of various systems for identifying claim sentences about COVID-19. We use ClaimBuster, which does not involve topic detection, as a low-precision high-recall baseline. We can see that the performance improves by leveraging a pre-trained NLI model as a zero-shot topic-filtering system to filter out claims related to the topics at hand. Furthermore, even humans have relatively low precision, demonstrating the difficulty in identifying sentences with claims. We hypothesize that this could due to the subjective nature of whether an assertion is a claim or just a statement. It was also evident when we measured the inter-annotator agreement for claim sentence detection, with a Cohen's kappa of 0.44 demonstrating moderate agreement. Nevertheless, the model performance is still considerably worse compared to human performance, showing potential for future work.

Model	P	R	F1
ClaimBuster	13.0	**86.5**	22.6
ClaimBuster + Zero-shot NLI	**21.8**	53.3	**30.9**
Human	52.7	70.0	60.1

Table 2: Performance (in %) of various systems for detecting claims related to COVID-19, given a news article as an input.

We also report the performance of the NLI model for topic classification. Given the gold-standard claim sentence, the classification accuracy is 46.6% for the zero-shot NLI model over these four topics.

6.2 Claim Object Detection

Setup: We use the development set to get the few-shot examples, by sampling[10] five examples for each topic. To account for the variance from sampling, we report numbers that are averaged over three runs. For language model sizes to be comparable, we use the Ada[11] version of GPT-3 and the base version of T5. We fine-tune the T5-base model for five epochs with a learning rate of 3e-5. The evaluation metric is string-match F1 similar to the one used in question answering (Rajpurkar et al., 2016).

Results and Analysis: Table 3 shows the F1 score for extracting the claim object relating to the topic. In zero-shot settings, we see that GPT-3 performs considerably better than T5, potentially benefiting from the larger corpus of data it was trained on. In the few-shot settings however, T5 is competitive with GPT-3, showing the promise of prompt-based fine-tuning, even with limited few-shot examples.

Approach	Model	Type	F1
Prompting	GPT-3	Zero-shot	15.2
Prompting	T5	Zero-shot	11.4
In-context learning	GPT-3	Few-Shot	**51.9**
Prompt-based fine-tuning	T5	Few-Shot	51.6
Human	-	-	67.7

Table 3: F1 score (in %) of different zero-shot and few-shot systems for the claim object detection sub-task.

6.3 Stance Detection

Setup: We use the same BART-large model trained for NLI as in Section 6.1. In the setting with access to the topic, it is taken from the gold-standard annotation.

Results and Analysis: We also consider a majority class baseline that always predicts *affirm* as the stance. Table 4 shows the performance of stance detection approaches. We can see that the NLI model with access to topic performs the best, with considerable improvement in performance for the *refute* class. Thus, access to additional background information helps here as the topic of the claim can be used to come up with a more relevant hypothesis, as is evident from Figure 4b.

6.4 Claim Span Detection

Results and Analysis: The evaluation measure in this setting is character-span F1. From Table 5, we see that the Debater claim boundary detection system considerably outperforms the attribution-based system. This could be because the former is trained on arguments, which are more similar to claims compared to statement-like attributions.

[9] http://huggingface.co/facebook/bart-large-mnli

[10] We will release the few-shot examples used in our sampling runs to make the results reproducible.

[11] https://blog.eleuther.ai/gpt3-model-sizes/

Model	Affirm F1	Refute F1	Acc.
Majority class	82.5	0.0	70.3
NLI (No topic)	89.1	68.0	83.8
NLI (With topic)	**91.1**	**78.8**	**87.5**
Human	97.0	84.2	94.9

Table 4: F1 score (in %) of both affirm and refute classes along with overall accuracy for stance detection. The zero-shot NLI system is shown separately based on access to topic while constructing the hypothesis.

Model	Prec.	Recall	F1
PolNeAR-Content	67.0	42.8	52.3
Debater Boundary Detection	**75.7**	**77.7**	**76.7**
Human	82.7	90.9	86.6

Table 5: Performance (in %) of different systems for identifying boundaries of the claim within a given claim sentence.

6.5 Claimer Detection

Setup: For the PolNeAR-Source system, the threshold for confidence score is tuned on the development set. The claim span output from the Debater boundary system is used marking the claim content within the context. For the SRL system, we leverage the parser[12] provided by AllenNLP (Gardner et al., 2018), which was trained on OntoNotes (Pradhan et al., 2013). The evaluation involves scores for the journalist (classification F1) and reported (string-match F1), along with overall F1.

Model	F1	Reported	Journalist
SRL	41.7	23.5	**67.2**
PolNeAR-Source	**42.3**	**25.5**	65.9
Human	85.8	81.3	88.9

Table 6: F1 score (in %) for claimer detection. Numbers for journalist and reported are shown separately, along with the overall F1.

Results and Analysis: From Table 6, we see that automatic models are considerably behind humans for claimer detection. While the performance is relatively better for the case of identifying whether the journalist is making the claim, models perform poorly for reported claims, which involves extracting the mentions of the claimer. For reported claims, we further understand the impact of location of the claimer mention within the news article. Table 7 shows the performance depending on whether the claimer is mentioned inside or outside the claim sentence. We see that these attribution models are able to handle claimer detection for reported claims only when the claimer mention is within the claim sentence. The need for cross-sentence reasoning for the claimer detection sub-task is evident from the low out-of-sentence F1 score for these sentence-level approaches.

[12] AllenNLP SRL Parser

Model	In-sentence	Out-of-sentence
SRL	35.8	2.4
PolNeAR-Source	**38.9**	**2.7**

Table 7: F1 score (in %) in terms of reported claims for extracting the claimer when it is present within (in-sentence) or outside (out-of-sentence) the claim sentence.

6.6 Remaining Challenges

Given the relatively low topic classification performance of the NLI model, the benchmark requires better zero-shot approaches for selecting claims relating to COVID-19. Furthermore, good topic classification performance is important as the claim topic is crucial to claim object detection and can also be leveraged for better stance detection.

The claimer detection subtask requires incorporating stronger cross-sentence reasoning to identify the claimer for when the mention is outside the claim sentence. This necessitates building attribution systems that are document-level. Furthermore, the same news article can have similar claims but from different claimers. To prevent misattribution in such cases, it would be beneficial to identify context within the news article that is relevant to the given claim, so as to remove noise from other related claims.

7 Conclusion

In this work, we proposed a new benchmark, NewsClaims, that extends the current claim detection task to extract more background attributes related to the claim. Our benchmark comprehensively evaluates multiple aspects of claim detection such as identifying topics, the stance, the claim span, the claim object, and the claimer in news articles from emerging scenarios such as the COVID-19 pandemic. We show that zero-shot and prompt-based few-shot approaches can achieve promising performance in such low resource scenarios, but still lag behind human-level performance, which presents opportunities for further research. Future work will explore is towards building a unified multi-task framework that can simultaneously identify multiple background attributes.

Ethics and Broader Impact

Misuse Potential The intended use of NEWSCLAIMS is to evaluate methodological work regarding our augmented definition of claim detection, motivated by mitigating the spread of misinformation and disinformation in news media. However, given NEWSCLAIMS is a smaller dataset over a set of hand-chosen topics, there is also potential for misuse. Specifically, NEWSCLAIMS is not intended to directly make conclusions regarding the journalism quality nor quantify disagreement regarding coverage of COVID-19 related topics. NEWSCLAIMS is not intended as a training dataset and a system using NEWSCLAIMS in this way should be carefully evaluated before being used to annotate a larger dataset aimed at deriving journalism-centric conclusions. As there has been continued controversy regarding media coverage of COVID-19, a bad faith or misinformed actor could produce artifacts that result in sensational, but potentially inaccurate, conclusions regarding COVID-19 claims in news media.

COVID-19 Specificity In this vein, NEWSCLAIMS exclusively consists of claims regarding COVID-19, intentionally chosen to sufficiently study a quickly emerging subject. However, performance on this dataset might likely not be representative of performance on a broader set of topics. In the future, we hope to mitigate these risks with a larger dataset that can more reliable to study these phenomena and produce conclusions about the underlying media content.

Environmental Impact We would also like to warn that the use of large-scale Transformers requires a lot of computations and the use of GPUs for training, which contributes to global warming (Strubell et al., 2019). This is a bit less of an issue in our case, as we do not train such models from scratch; rather, we mainly use them in zero-shot and few-shot settings, and the ones we fine-tune are on relatively small datasets. All our experiments were run on a single 16GB V100.

References

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, et al. 2020. Fighting the covid-19 infodemic: modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. *arXiv preprint arXiv:2005.00033*.

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A benchmark dataset of check-worthy factual claims. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 821–829.

Roy Bar-Haim, Yoav Kantor, Elad Venezian, Yoav Katz, and Noam Slonim. 2021. Project Debater APIs: Decomposing the AI grand challenge. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 267–274, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, pages 3–5.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22.

David K Elson and Kathleen R McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Association for Computational Linguistics (ACL)*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018.

Allennlp: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6.

Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, Varna, Bulgaria. INCOMA Ltd.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. A survey on stance detection for mis-and disinformation identification. *arXiv preprint arXiv:2103.00242*.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017. Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. ClaimRank: Detecting check-worthy claims in Arabic and English. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, New Orleans, Louisiana. Association for Computational Linguistics.

Ye Jiang, Xingyi Song, Carolina Scarton, Ahmet Aker, and Kalina Bontcheva. 2021. Categorising fine-to-coarse grained misinformation: An empirical study of covid-19 infodemic. *arXiv preprint arXiv:2106.11702*.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500.

Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081.

Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. 2017. Unsupervised corpus–wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining*, pages 79–84.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Salman Bin Naeem and Rubina Bhatti. 2020. The covid-19 ‘infodemic’: a new front for information professionals. *Health Information & Libraries Journal*, 37(3):233–239.

Edward Newell, Drew Margolin, and Derek Ruths. 2018. An attribution relations corpus for political news. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107.

Silvia Pareti. 2016a. Parc 3.0: A corpus of attribution relations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 3914–3920.

Silvia Pareti. 2016b. PARC 3.0: A corpus of attribution relations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 3914–3920, Portorož, Slovenia. European Language Resources Association (ELRA).

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

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

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*, pages 2383–2392.

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 (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352.

Shaden Shaar, Maram Hasanain, Bayan Hamdan, Zien Sheikh Ali, Fatima Haouari, Alex Nikolov, Mücahid Kutlu, Yavuz Selim Kartal, Firoj Alam, Giovanni Da San Martino, et al. 2021. Overview of the clef-2021 checkthat! lab task 1 on check-worthiness estimation in tweets and political debates. In *CLEF (Working Notes)*.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pages 3645–3650, Florence, Italy.

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.

Adina Williams, Nikita Nangia, and Samuel 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, Volume 1 (Long Papers)*, pages 1112–1122.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3914–3923.

A Appendix

A.1 Annotation Interface

In this section, we list the annotation guidelines and provide screenshots of the interface for both phases of annotation. Phase 1 of annotation involves identifying sentences which contain claims relating to set of pre-defined topics about COVID-19. Phase 2 consists of annotating the background attributes such as claimer, claimer's stance, claim object and the claimer span for each of the claims identified in phase 1. Figure 6 and 7 show screenshots of the annotation interface for phase 1 and 2 respectively. Below are some guidelines which we provide for detecting the claim sentences:

- The highlighted sentence should be considered individually when deciding whether it contains a claim. The sentences around it are shown to provide context.
- Claims are usually statements made without presenting evidence or proof, and usually require further evidence to verify them. Sentences that just assert evidences or present facts should not be considered as claims.
- The claim sentences usually should also mention the object relating to the topic, i.e which animal the virus came from, what conditions can transmit the virus, what can cure the virus or what can protect from the virus.
- Only those claims should be considered for which these topics can be directly inferred without any need for additional knowledge.
- Sentences that contain both claims as well as refute statements should be considered. For e.g. A sentence that contains a statement that something cannot cure the coronavirus should be considered as containing a claim relating to the topic: Cure for the virus.

A.2 SRL cue words

Here, we list the different cue words that we use to match against the verb predicates from the SRL parser. These are categorized as affirming and refuting cue words, which are shown in tables 8 and 9 respectively.

A.3 GPT-3 prompt

In this section, we share more details of our approach for prompting GPT-3 for the claim object detection. In the in-context learning setting, we choose four examples from each topic as the few-shot examples. These labeled examples are then added to the context that is fed as input to GPT-3. The test example is added at the end of the context, in the form of a prompt, with the claim object to be generated by the system. Figure 5 shows an example input along with the prompt.

accuse, affirm, allege, announce, argue
assert, aver, avouch, avow, blame
broadcast, claim, comment, confirm, contend
credit, declare, defend, describe, disclose
discuss, express, find, hint, imply
insinuate, insist, intimate, maintain, proclaim
profess, publish, purport, reaffirm, reassert
remark, repeat, report, restate, reveal
say, state, suggest, tell, write

Table 8: Cue words corresponding to affirming a claim.

challenge, controvert, contradict,disagree
discredit, dispute, deny, disavow, discount
protest, purport, reaffirm, question, repudidate
reject, repudiate, rebut, suppress, disaffirm

Table 9: Cue words corresponding to refuting a claim.

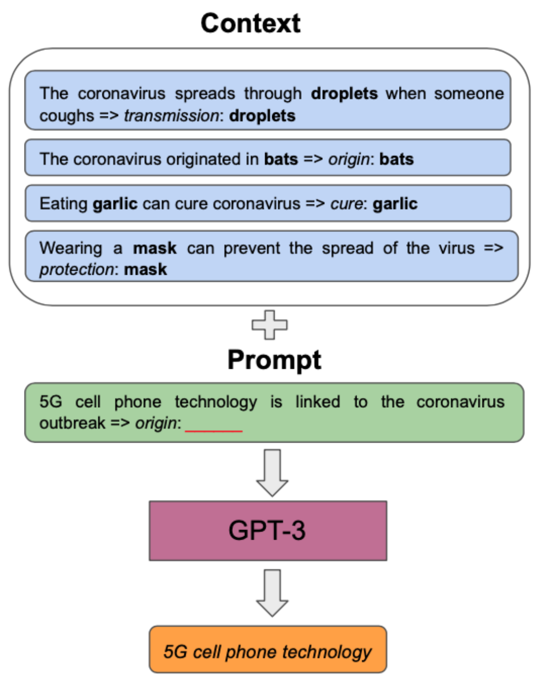
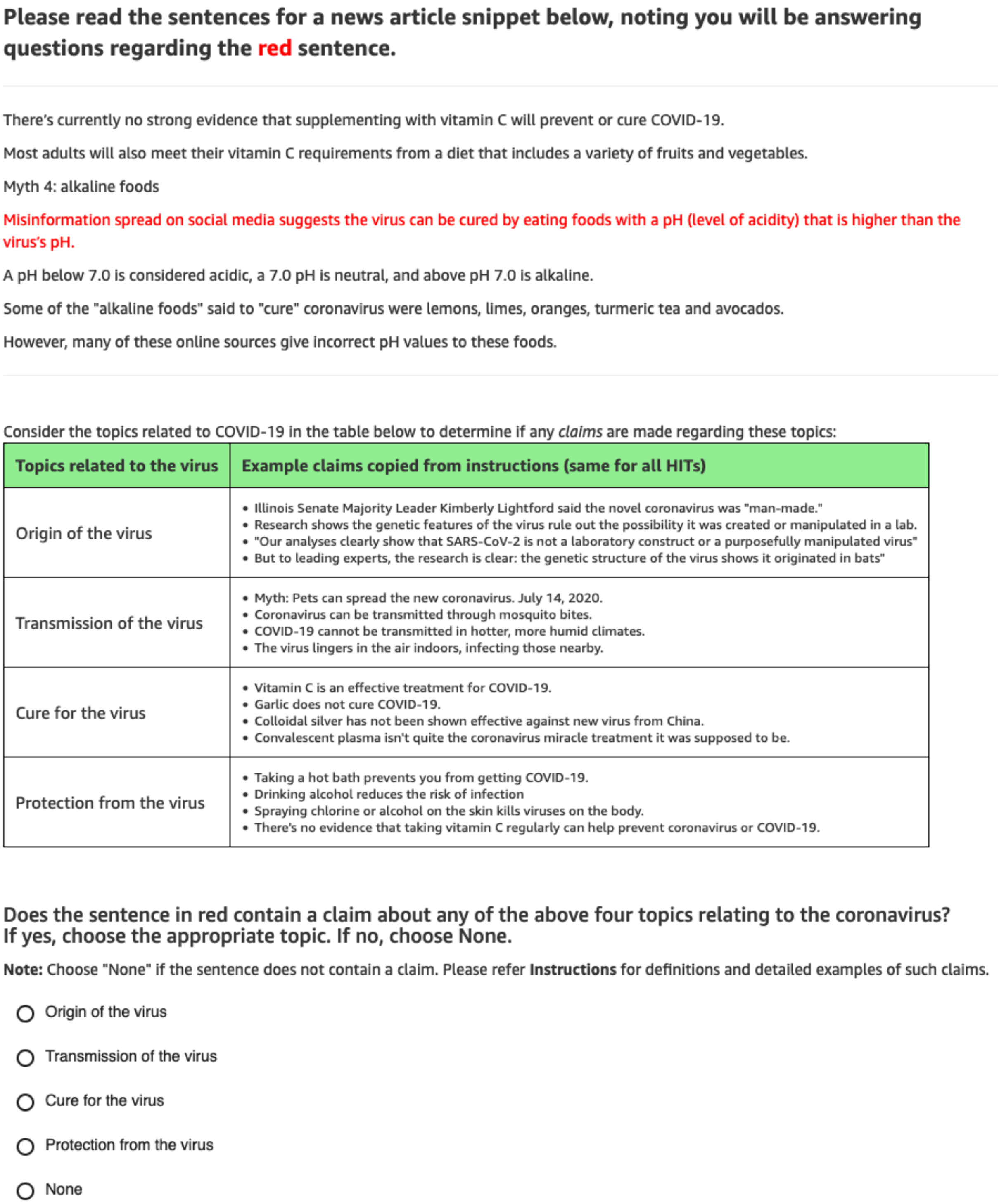


Figure 5: Figure showing the claim object detection sub-task input for GPT-3, with the few-shot labeled examples in context and the test example in the form of a prompt.



Please read the sentences for a news article snippet below, noting you will be answering questions regarding the red sentence.

There's currently no strong evidence that supplementing with vitamin C will prevent or cure COVID-19.

Most adults will also meet their vitamin C requirements from a diet that includes a variety of fruits and vegetables.

Myth 4: alkaline foods

Misinformation spread on social media suggests the virus can be cured by eating foods with a pH (level of acidity) that is higher than the virus's pH.

A pH below 7.0 is considered acidic, a 7.0 pH is neutral, and above pH 7.0 is alkaline.

Some of the "alkaline foods" said to "cure" coronavirus were lemons, limes, oranges, turmeric tea and avocados.

However, many of these online sources give incorrect pH values to these foods.

Consider the topics related to COVID-19 in the table below to determine if any *claims* are made regarding these topics:

Topics related to the virus	Example claims copied from instructions (same for all HITs)
Origin of the virus	• Illinois Senate Majority Leader Kimberly Lightford said the novel coronavirus was "man-made." • Research shows the genetic features of the virus rule out the possibility it was created or manipulated in a lab. • "Our analyses clearly show that SARS-CoV-2 is not a laboratory construct or a purposefully manipulated virus" • But to leading experts, the research is clear: the genetic structure of the virus shows it originated in bats"
Transmission of the virus	• Myth: Pets can spread the new coronavirus. July 14, 2020. • Coronavirus can be transmitted through mosquito bites. • COVID-19 cannot be transmitted in hotter, more humid climates. • The virus lingers in the air indoors, infecting those nearby.
Cure for the virus	• Vitamin C is an effective treatment for COVID-19. • Garlic does not cure COVID-19. • Colloidal silver has not been shown effective against new virus from China. • Convalescent plasma isn't quite the coronavirus miracle treatment it was supposed to be.
Protection from the virus	• Taking a hot bath prevents you from getting COVID-19. • Drinking alcohol reduces the risk of infection • Spraying chlorine or alcohol on the skin kills viruses on the body. • There's no evidence that taking vitamin C regularly can help prevent coronavirus or COVID-19.

Does the sentence in red contain a claim about any of the above four topics relating to the coronavirus? If yes, choose the appropriate topic. If no, choose None.

Note: Choose "None" if the sentence does not contain a claim. Please refer **Instructions** for definitions and detailed examples of such claims.

- Origin of the virus
- Transmission of the virus
- Cure for the virus
- Protection from the virus
- None

Figure 6: Screenshot of the phase 1 annotation interface.

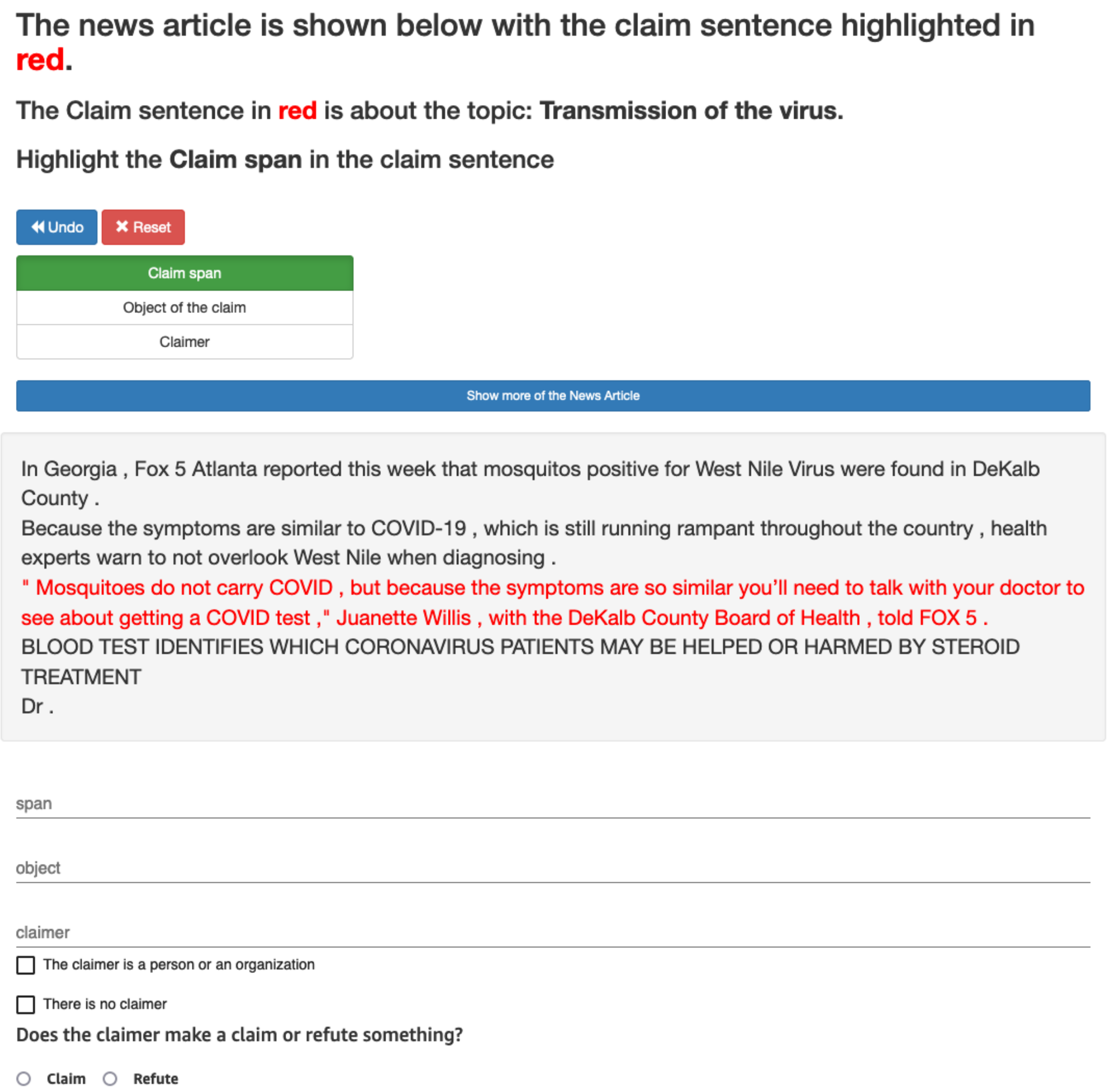


Figure 7: Screenshot of the phase 2 annotation interface.

News Text: Another recent paper in Nature Medicine underscores that point. *"By comparing the available genome sequence data for known coronavirus strains, we can firmly determine that SARS-CoV-2 originated through natural processes," Kristian Andersen, PhD, an associate professor of immunology and microbiology at Scripps Research and corresponding author on the paper, said in a statement.* Andersen and colleagues' research implicates bats and possibly pangolins.

Topic: Origin of the virus
Stance: Affirm
Claim Object: natural processes
Claimer: Kristian Andersen

News Text: MYTH: Antibiotics are effective at treating coronavirus. *"Antibiotics do not treat viruses of any kind, including coronaviruses," Dr. Kenney states unequivocally*. Antibiotics target bacteria, not viruses.

Topic: Cure of the virus
Stance: Refute
Claim Object: Antibiotics
Claimer: Dr. Kenney

News Text: Does chlorine in pool water inactivate the virus? *"The good news is that the average amount of chlorine that's in a pool is going to kill the virus," Lavin says.* Assuming that your pool is properly maintained, the disinfecting chemicals in the water should be enough to render the virus inactive.

Topic: Protection from the virus
Stance: Affirm
Claim Object: chlorine
Claimer: Lavin

News Text: Airborne Transmission in Tight Spaces. *Medical professionals from the preeminent organizations on public health Centers for Disease Control and Prevention (CDC) and the World Health Organization have started changing their stance that COVID-19 is airborne.* This is important news.

Topic: Transmission of the virus
Stance: Affirm
Claim Object: air
Claimer: Medical professionals

News Text: That the virus has natural origins is also apparent from its molecular structure. *Scientists writing in Nature Medicine journal on March 17 made clear that "all notable SARS-CoV-2 features" were also observed "in related coronaviruses in nature" and that therefore "we do not believe that any type of laboratory-based scenario is plausible."* As Josie Golding, epidemics lead at the UK-based Wellcome Trust, pointed out: The findings "are crucially important to bring an evidence-based view to the rumors that have been circulating about the origins of the virus."

Topic: Origin of the virus
Stance: Refute
Claim Object: laboratory
Claimer: Scientists

News Text: Gilead is working with the U.S. government on the logistics of remdesivir distribution and will provide more information when the company begins shipping the drug under the EUA. *"This EUA opens the way for us to provide emergency use of remdesivir to more patients with severe symptoms of COVID-19," said Daniel O'Day, Chairman and Chief Executive Officer of Gilead Sciences*. "We will continue to work with partners across the globe to increase our supply of remdesivir while advancing our ongoing clinical trials to supplement our understanding of the drug's profile.

Topic: Cure of the virus
Stance: Affirm
Claim Object: remdesivir
Claimer: Daniel O'Day

News Text: We reached out to him for a comment, but we haven't heard back. *On its website, the World Health Organization says that, while it recommends eating plenty of fruits and vegetables to stay healthy, there is no scientific evidence that lemon treats or prevents COVID-19 infection.* To prevent coronavirus infection, officials advise people to regularly wash their hands, avoid touching their face, disinfect surfaces in their homes daily and avoid people who are sick.

Topic: Protection from the virus
Stance: Refute
Claim Object: lemon
Claimer: World Health Organization

News Text: If you are not staying in the same spot, like moving through a grocery store or walking, then your rate of infection decreases. *Doctors are concerned with the dosage of droplets that leads to infection.* As of the publication of this blog, doctors have not specified a dosage rate required for infection.

Topic: Transmission of the virus
Stance: Affirm
Claim Object: droplets
Claimer: Droplets

Figure 8: Some examples from the NEWSCLAIMS benchmark.