Check-COVID: Fact-Checking COVID-19 News Claims with Scientific Evidence

Gengyu Wang[1] **Kate Harwood** [1] **Lawrence Chillrud** [2]
Amith Ananthram [1] **Melanie Subbiah**[1] **Kathleen McKeown**[1]
[1]Columbia University [2]Northwestern University
{gengyu.wang, k.r.harwood}@columbia.edu, chili@u.northwestern.edu
{amith.ananthram, m.subbiah}@columbia.edu, kathy@cs.columbia.edu

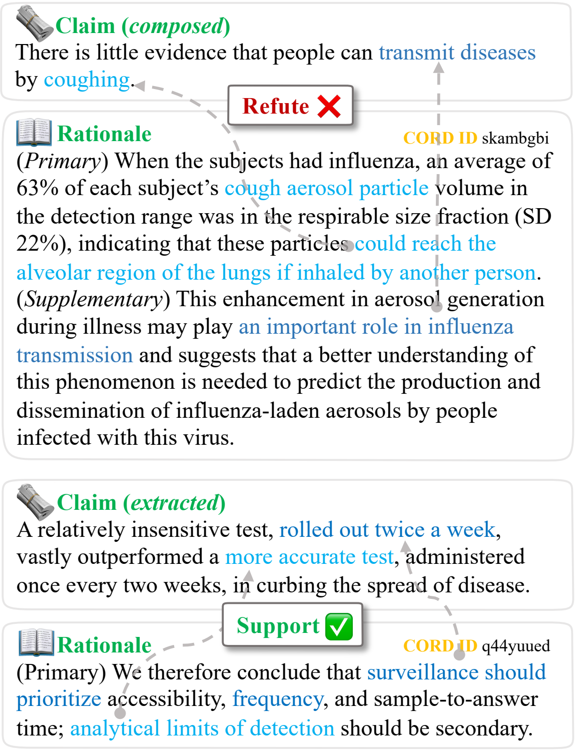
Abstract

We present a new fact-checking benchmark, Check-COVID, that requires systems to verify claims about COVID-19 from news using evidence from scientific articles. This approach to fact-checking is particularly challenging as it requires checking internet text written in everyday language against evidence from journal articles written in formal academic language. Check-COVID contains $1,504$ expert-annotated news claims about the coronavirus paired with sentence-level evidence from scientific journal articles and veracity labels. It includes both *extracted* (journalist-written) and *composed* (annotator-written) claims. Experiments using both a fact-checking specific system and GPT-3.5, which respectively achieve F1 scores of 76.99 and 69.90 on this task, reveal the difficulty of automatically fact-checking both claim types and the importance of in-domain data for good performance. Our data and models are released publicly at https://github.com/posuer/Check-COVID.

1 Introduction

Throughout the COVID-19 pandemic, misinformation on the internet has proven to be exceptionally dangerous, undercutting containment efforts by public health officials around the world (Mheidly and Fares, 2020). In future pandemics, the ability to automatically detect and debunk such misinformation has the potential to save many lives. However, an automated system cannot rely solely on surface forms or linguistic features to identify misinformation (Pérez-Rosas et al., 2017; Rashkin et al., 2017). Such a system must be capable of checking claims written in everyday language against jargon-laden evidence from an evolving set of scientific articles.

In this work, we formalize this challenge as a new fact-checking benchmark that requires verifying claims about COVID-19 from news using evidence from scientific articles. This task is particularly challenging as it requires grounding everyday vernacular in formal academic language. Consider the logical inferences required to correctly label the first example presented in Figure 1 as refuted by the evidence. An automated system must recognize that *coughing* produces *aerosol particles* that may *reach the alveolar region of the lungs of another person*, which implies infection of that person. A successful model must be able to align such everyday language with scientific terminology while leveraging commonsense knowledge.



Claim (*composed*)
There is little evidence that people can transmit diseases by coughing.

Rationale CORD ID skambgbi
(*Primary*) When the subjects had influenza, an average of 63% of each subject's cough aerosol particle volume in the detection range was in the respirable size fraction (SD 22%), indicating that these particles could reach the alveolar region of the lungs if inhaled by another person. (*Supplementary*) This enhancement in aerosol generation during illness may play an important role in influenza transmission and suggests that a better understanding of this phenomenon is needed to predict the production and dissemination of influenza-laden aerosols by people infected with this virus.

Claim (*extracted*)
A relatively insensitive test, rolled out twice a week, vastly outperformed a more accurate test, administered once every two weeks, in curbing the spread of disease.

Rationale CORD ID q44yuued
(Primary) We therefore conclude that surveillance should prioritize accessibility, frequency, and sample-to-answer time; analytical limits of detection should be secondary.

Figure 1: Check-COVID examples. *Composed* claims are annotator-written based on assertions in news articles, whereas *extracted* claims are copied verbatim.

To facilitate research on this task, we introduce Check-COVID, a fact-checking dataset of $1,504$ claims about COVID-19, each of which is paired with sentence-level scientific evidence for its veracity label. These claims are drawn from news

Findings of the Association for Computational Linguistics: ACL 2023, pages 14114–14127
July 9-14, 2023 ©2023 Association for Computational Linguistics

articles, and the evidence is selected by human annotators from a large COVID-19 biomedical literature corpus, CORD-19 (Wang et al., 2020). The dataset includes 322 *extracted* claims, with wording drawn directly from the article, and $1,182$ *composed* claims, re-worded by trained annotators. Other fact-checking datasets have focused solely on one type of claim or the other. However, both types are important since people often include direct quotes or their own re-wording of claims when sharing articles online. Thus our dataset allows us to explore how models handle both these real-world claim types within the same domain. To build this dataset, we adapt SciFact's annotation method (Wadden et al., 2020) to news, leveraging naturally occurring citations in articles. Such citations usually contain claims with links to supporting journal articles. We first collect these claims and the abstracts of the referenced scientific articles and then manually identify the sentences in each abstract that provide claim-specific evidence.

To establish a baseline score, we adapt the fact-checking system presented in DeYoung et al. (2019) and Wadden et al. (2020), which uses an abstract retriever, rationale selector, and label predictor. To augment our training data, we experiment with existing fact-checking datasets that use a similar task formulation but contain claims from Wikipedia articles (FEVER (Thorne et al., 2018)), scientific journal articles (SciFact (Wadden et al., 2020)), or Reddit (COVID-Fact (Saakyan et al., 2021)) in addition to our corpus, Check-COVID. We train the models on different combinations of these four datasets, compare their performance, and select the best performing models to generate the baseline score. Experimental results make clear the difficulty of adapting existing corpora to this new benchmark - their inclusion yields only small improvements in performance. Moreover, an evaluation of GPT-3.5 (Brown et al., 2020) reveals limitations of in-context learning, especially in providing human-aligned evidence for its veracity labels.

Our contributions are: 1) we introduce a novel dataset (Check-COVID) as a benchmark for the challenging problem of fact-checking COVID-19 claims from news against evidence from scientific journal articles; 2) we evaluate how well models can adapt to both *extracted* and *composed* claims within the same domain; and 3) we present a strong baseline for Check-COVID to facilitate future work. Our data and models will be released publicly under the MIT License at `https://github.com/posuer/Check-COVID`.

2 Related Work

There has been tremendous interest in developing fact-checking benchmarks for COVID-19, however many of these datasets only contain claims (Shah et al. (2022)) or claims with veracity labels and no evidence (Li et al. (2022); Elhadad et al. (2020); Shahi and Nandini (2020); Cui and Lee (2020); Vijjali et al. (2020); Alam et al. (2020)). Others primarily source claims from social media (Saakyan et al. (2021); Mohr et al. (2022); Sundriyal et al. (2022)) or scientific texts (Wadden et al. (2020)). Some datasets, like ours, source claims from colloquial news sources (or other online sources) (Lee et al. (2020); Lee et al. (2021); Sarrouti et al. (2021)) but of these Lee et al. (2020) and Lee et al. (2021) have just a few hundred examples, a challenge for use in training deep learning systems.

Sarrouti et al. (2021) is the most similar to our dataset, however, our evidence comes from citations in the news articles the claims are drawn from, whereas in Sarrouti et al. (2021) the evidence is retrieved using the extracted claims as search queries. Thus, our claim-evidence pairings are closer to those the average reader encounters online.

Additionally we include a third veracity label (NOTENOUGHINFO) which some of the above datasets - and many general fact-checking corpora - do not. In the wild, a system will not always be able to fact-check a claim. Thus, modeling these cases is critical to real-world performance.

3 Check-COVID Dataset

We introduce Check-COVID, a dataset of $1,504$ claims drawn from news paired with sentence-level evidence from scientific journal articles and accompanying veracity labels: $\{$SUPPORT, REFUTE, NOTENOUGHINFO$\}$. The number of examples across the three labels is balanced $(505, 504, 495)$. The claims are categorized into *composed* or *extracted* based on whether they are written by our annotators or drawn from news articles. We also provide a corpus containing the abstracts of the journal articles from which the evidence is drawn, a subset of the CORD-19 corpus. We randomly split the dataset into three balanced subsets with no overlapping abstracts: train (70%), dev (15%) and test (15%).

3.1 Data Source

Citances (Nakov et al., 2004) in news are spans of text that contain assertions about findings from scientific journals with accompanying citations to the articles where the supporting evidence can be found. We build a crawler to automatically detect citations in news and collect sentences surrounding the citance, the abstract of the cited journal article, and the articles' URLs. To ensure the cited information is scientific, the crawler ignores citances which do not cite journal articles in CORD-19, a trustworthy corpus of scientific papers on coronavirus research. We restrict our corpus to a set of well-regarded news websites[1] to increase the likelihood that claims are paired with relevant scientific evidence (fake news, in contrast, regularly contains deceptive citances). Our manual annotation and claim negation process is then a second check to ensure that cited evidence is appropriate.

3.2 Claims in Check-COVID

In Check-COVID, a claim is an atomic factual statement describing one aspect of a scientific entity or process related to COVID-19 (such that it can be fact-checked against primary research). For example, "*Cloth masks offer significantly less protection against infection than medical masks.*" Opinions or facts that do not require scientific proof are not considered valid claims (e.g., "*The government published a policy requiring masks in public spaces*").

Composed and Extracted Check-COVID contains both *composed* and *extracted* claims. To generate *composed* claims, we present annotators with news paragraphs that contain a citance. Annotators are asked to write a claim based on the information in the paragraph. We require that *composed* claims be understandable without extra context. For *extracted* claims, we detect claims from citance-containing news paragraphs using a claim detection model (Barrón-Cedeno et al., 2019) trained on ClaimBuster (Arslan et al., 2020), then we present them to annotators to decide whether they satisfy our definition of a claim. In Figure 1, we present two claims, one *composed* and one *extracted*. *Composed* claims are usually shorter, simpler, and more similar to the kind of claims that the general public might write online about an article or submit to a fact-checking system. In contrast, *extracted* claims are retrieved directly from news articles, often use more complicated wording, and are necessary to train systems that fact-check news directly. Our experiments demonstrate that models trained on one category cannot simply be adapted to the other. In Check-COVID, the ratio between *composed* claims and *extracted* claims is $3.67 : 1$. We explore differences between them in §**Dataset Statistics** and Table 1.

Claim Negation To create examples that are refuted by the cited abstracts, we manually negate the supported claims. Negation procedures can introduce bias in a dataset which can allow a model to "cheat" on a task. For example, a model can learn to associate the word "not" with the REFUTE label (Schuster et al., 2019). To mitigate these effects, we request that annotators avoid using negators like "does not", "cannot", "no", etc. Instead, we ask them to negate claims by changing them more fundamentally, for example, by changing *Cloth masks offer much less protection against infection than medical masks* to *Cloth masks are just as effective at preventing infection as medical masks*. As negations involve minimal changes to the original claim style, negated claims retain the type designation (*composed* or *extracted*) of their originals.

3.3 Annotation Procedure

On a web-based annotation interface (see Appendix for screenshots), annotators are shown a news paragraph or a selection of automatically detected claims together with one of the paragraph's cited abstracts. Annotators start by either selecting a valid claim from those provided or writing a *composed* claim from the full news paragraph. In either case, they then write a second claim that is the negation of the first. For each claim, annotators identify whether the claim is "SUPPORTED" or "REFUTED" by the abstract and select rationales from the abstract as justification. A rationale is a minimal collection of sentences sufficient to justify a label of SUPPORT or REFUTE in relation to a claim.

In fact-checking, often there is not adequate evidence to confidently certify or debunk a claim, so we also introduce a third label, NOTENOUGHINFO. To create a NOTENOUGHINFO example, we randomly sample from the *composed* claims and pair the sampled claim with a sentence from its corresponding abstract that was not chosen by an annotator as evidence for that claim. By choosing non-evidence sentences from the cited abstract, we select sentences that exhibit both lexical and topic overlap

[1] Including The New York Times, The Washington Post, and BBC. The full list is in the Appendix A.

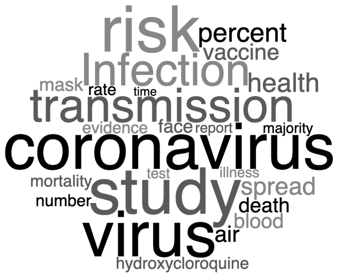


Figure 2: Common words in Check-COVID claims.

with the claim without providing evidence for it, resulting in difficult NOTENOUGHINFO examples.

3.4 Quality Control

We employ four graduate students with a background in NLP and four graduate students studying life sciences as annotators through a mailing list of the education institution. We paid them the minimum hourly rate required by our locality and obtained signed informed consent forms that explain how the collected data would be used. We obtained approval (ethics review) for this study from the Institutional Review Board of the authors' institution. As our data is collected or derived from publicly available sources, it is not subject to any anonymization practices. We removed any identifying information from the annotations. All annotators watch a video, read an instruction guide and produce practice annotations to become familiar with the task. We include the instructions in the Appendix B and we make the video available online[2]. As annotators need to write their own versions of claims, it is difficult to calculate an agreement score for their annotations. Therefore, to control quality, each annotator is assigned to an NLP graduate student who reviews each annotation, provides revision suggestions, and monitors revisions until the annotation meets the requirements described above. Approved annotators then review each other's submissions. At the same time, expert annotators continue monitoring new annotations for quality and provide feedback when necessary. As a final check, all submitted claims are proofread by the authors. We believe such adjudication, though laborious, is of renewed importance in our field, providing high quality benchmarks that accurately measure model performance (Pustejovsky and Stubbs, 2012).

	REFUTE	NOTENINFO	SUPPORT
Composed			
Number of claims	419	392	371
Avg. # words per claim	16.54	16.66	16.72
% claims containing *not/no*	0.32	0.24	0.10
Avg. # sentences per rationale	1.45	1.00	1.46
Avg. # words per evidence sentences	33.32	28.64	33.20
Extracted			
Number of claims	85	103	134
Avg. # words per claim	29.95	29.35	29.03
% claims containing *not/no*	0.35	0.19	0.13
Avg. # sentences per rationale	1.54	1.0	1.57
Avg. # words per evidence sentence	33.04	26.94	32.98

Table 1: Check-COVID summary statistics.

3.5 Dataset Statistics

Check-COVID covers a diverse set of topics as demonstrated by the variety of keywords in claims shown in Figure 2. Table 1 presents summary statistics for our corpus. We can see that *extracted* claims are twice as long as *composed* claims, suggesting that *composed* claims use simpler wording. This linguistic diversity, along with the differences in system performance in Section 6, is further evidence for the importance of including both claim types as part of a fact-checking benchmark.

Since we annotate REFUTE claims by manually negating SUPPORT claims, the "not/no" in REFUTE claims are either from their source SUPPORT claims or introduced by the annotation process. While we request that annotators avoid using "not/no" during negation, the percentage of claims containing "not/no" shows that more REFUTE claims contain "not/no" than SUPPORT claims. However, our results for label prediction in Table 3 show that this correlation does not result in a clear pattern of REFUTE claims being easier to check than SUPPORT claims.

4 Check-COVID Task and Challenges

Given a claim about COVID-19, either written by an annotator or drawn from a news article, the task is to retrieve related scientific abstracts from CORD-19, choose the most relevant evidence sentences from those abstracts and decide whether that evidence supports, refutes, or does not provide enough information about the claim. This formulation matches the SciFact task format. The claims in our corpus pose many unique challenges:

Medical Knowledge and Terminology Many examples in our corpus require knowledge of medicine and its terminology to perform successful fact-checking. To fact-check the first example in Figure 1, a model must recognize that the *alveolar*

[2] https://youtu.be/mEOcouML9oA

region is *where blood exchanges oxygen and carbon dioxide*[3]. This implies that here, the virus can enter someone's bloodstream and infect them.

Temporal Reasoning The examples in Check-COVID also require understanding and comparing temporal durations, frequencies and ordering. To fact-check the second claim in Figure 1; a model needs to understand that testing *twice a week* is more frequent than *once every two weeks*.

Numeric Values Most examples with numeric values require numeric comparison and number-word translation. For example, given the claim *communities of color were disproportionately affected by COVID-19* and its evidence *cumulative incidence was... higher among Hispanic/Latino (29.2%)... than non-Hispanic white adults (8.1%, p<.0001)*, a model must compare *29.2%* and *8.1%* and deem this difference *disproportionate*.

We closely scrutinize the prevalence and impact of these challenges. With regard to medical terminology and knowledge, all examples in our dataset pose this challenge, as our task requires fact-checkers to understand and extract evidence from medical journal articles. Regarding temporal reasoning and numeric reasoning, there are 147 and 93 claims across a total of 450 dev and test set examples respectively that require such reasoning to be fact-checked. Although we will delve deeper into the fact-checking system and related experiments in forthcoming sections, our analysis indicates a decline in the system's performance on this specific subset. Notably, the macro F1 of the composed test set processed through the Vespa pipeline fell from 63.34 to 29.80 for the temporal subset, and to 33.87 for the numeric subset.

5 Baseline System

Our fact-checking pipeline is adapted from the systems presented in DeYoung et al. (2019), Wadden et al. (2020), and Wang et al. (2021), whose use is consistent with their intended use. It consists of three modules, trained in a supervised fashion on different combinations of existing fact-checking corpora and Check-COVID. We describe the individual components of this baseline system below.

Abstract Retrieval We query for COVID-19 scientific literature through a CORD-19 search engine, *Vespa*[4]. Given a claim, *Vespa* retrieves relevant abstracts from CORD-19 and ranks them with the BM25 scoring function (Robertson et al., 1995). By integrating *Vespa* instead of using a retrieval model with a fixed corpus, we enable the pipeline to retrieve evidence from a corpus of COVID-19 scientific articles that could be continually updated. The three abstracts with the highest BM25 scores are passed to the next step of *Rationale Selection*.

Rationale Selection We fine-tune a RoBERTa (Liu et al., 2019) model with Transformers (Wolf et al., 2020) to predict whether each sentence e_i from the selected abstracts is relevant to the given claim c, where i refers to the index within the abstract. We encode each sequence $[c\ SEP\ e_i]$ and then feed the final $[CLS]$ representation to a linear classification layer to predict a binary relevance label. Sentences that score over 0.5 are passed as a (possibly empty) set to *Label Prediction*.

Label Prediction We fine-tune another RoBERTa model to label each selected rationale e_i as SUPPORT, REFUTE, or NOTENOUGHINFO. Given a claim c and evidence e_i, we encode the sequence $[c\ SEP\ e_i]$ with RoBERTa and then feed the final $[CLS]$ representation to a linear classification layer. We optimize a cross-entropy loss over our 3 classes. For claims with a multi-sentence rationale, e_i is the concatenation of the sentences.

6 Experiments

To establish a baseline score for Check-COVID, we first evaluate the dev-set performance of our *Rationale Selection* and *Label Prediction* modules. We then choose the best-performing model for each to build the full end-to-end pipeline system which we evaluate on the test split. Given the material differences between *composed* and *extracted* claims, we evaluate training on each claim type separately and in tandem. As Vespa could return many relevant abstracts from CORD-19 beyond the ones selected for our corpus, we omit evaluation of the *Vespa*-based *Abstract Retrieval* module by itself.

Training Datasets To build our *Rationale Selection* and *Label Prediction* modules, we fine-tune RoBERTa-Large models sequentially on FEVER and/or SciFact and then on Check-COVID. We note that though FEVER and SciFact contain claims and evidence unrelated to COVID-19, their inclusion allows us to evaluate the transfer potential of additional task-related data from different domains. For *Label Prediction*, we also explore fine-tuning on COVID-Fact (which we cannot use for *Ratio-*

[3] from NCI: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/alveoli

[4] https://cord19.vespa.ai/

	Sentence Level						Rationale Level	
Rationale Selection	**Standard**			**Global Recall-Focused**			**Strict**	**Intersection**
	Precision	Recall	F1	Precision	Recall	F1	Accuracy	Accuracy
Evaluated on **Composed**								
** Train on composed*								
Check-COVID	53.20	34.72	39.95 (6.31)	44.80	28.24	32.91 (2.34)	19.81 (1.89)	44.65 (16.05)
FEVER + Check-COVID	47.56	53.47	50.25 (1.03)	34.82	39.12	**36.78**†(2.26)	22.96 (3.93)	**69.81** (2.50)
SciFact + Check-COVID	52.67	45.14	48.26 (1.66)	38.28	32.87	35.12 (2.18)	24.53 (5.25)	60.06 (5.76)
FEVER + SciFact + Check-COVID	54.38	49.07	**51.23** (1.83)	38.90	34.95	36.56 (1.14)	**29.25** (2.50)	64.78 (7.57)
** Train on composed **and** extracted together*								
Check-COVID	47.75	47.69	47.27 (0.96)	38.69	38.66	**38.31**†(0.54)	23.27 (6.82)	62.26 (5.74)
FEVER + Check-COVID	50.87	53.01	51.70 (1.06)	37.14	38.89	37.83 (2.87)	26.42 (1.63)	68.87 (5.74)
SciFact + Check-COVID	53.80	50.46	52.05 (1.74)	38.28	35.88	37.02 (2.70)	**26.73** (1.44)	66.35 (1.44)
FEVER + SciFact + Check-COVID	47.76	57.87	**52.18** (0.40)	33.99	41.20	37.14 (0.88)	24.84 (3.31)	**75.16** (5.20)
GPT-3.5 (text-davinci-003)	51.35	39.58	44.71	42.34	32.64	36.86	16.98	49.06
Evaluated on **Extracted**								
** Train on extracted*								
Check-COVID	72.49	46.26	56.39 (1.37)	36.05	23.13	28.13 (4.00)	25.25 (3.78)	62.63 (1.43)
FEVER + Check-COVID	73.16	59.18	**65.40** (0.90)	49.69	40.14	**44.39**†(3.05)	**41.41** (1.75)	**79.80** (1.75)
SciFact + Check-COVID	70.30	50.34	58.62 (6.05)	39.93	28.57	33.28 (4.06)	31.31 (3.50)	67.68 (8.75)
FEVER + SciFact + Check-COVID	65.56	53.06	58.21 (1.76)	46.16	37.41	41.01 (2.47)	31.31 (3.50)	73.74 (4.63)
** Train on composed **and** extracted together*								
Check-COVID	56.44	57.82	56.52 (1.08)	36.98	38.78	37.46 (5.35)	24.24 (3.03)	71.72 (4.63)
FEVER + Check-COVID	69.43	55.10	**61.38** (2.16)	39.28	31.29	34.80 (2.29)	**36.36** (3.03)	75.76 (3.03)
SciFact + Check-COVID	57.31	52.38	54.42 (3.97)	31.98	29.25	30.38 (2.03)	32.32 (3.50)	71.72 (12.25)
FEVER + SciFact + Check-COVID	58.45	63.95	60.90 (1.27)	38.92	42.86	**40.67**†(2.24)	30.30 (5.25)	**80.81** (1.75)
GPT-3.5 (text-davinci-003)	58.33	42.86	49.41	36.11	26.53	30.59	18.18	57.58

Table 2: Results (avg. of 3 seeds, std in parentheses) for *Rationale Selection* with different configurations of *composed* and *extracted* claims from Check-COVID and with or without FEVER and SciFact for training data. Metric details in Section 6.1.

nale Selection as it does not contain full abstracts). As we only use these corpora for research, it is consistent with their intended use.

Training Setting While training the *Rationale Selection* module, we pair each claim with its labeled evidence sentences to produce positive examples from FEVER, SciFact, or Check-COVID; we create negative examples by pairing claims and non-evidence sentences from the same abstracts that contained the labeled evidence. Since NOTENOUGHINFO claims are not paired with relevant evidence, we only use SUPPORT/REFUTE data points for training this module. For the *Label Prediction* module, we train on each claim and the concatenation of its gold-label evidence sentences.

Hyperparameter Settings For the rationale and label prediction modules, we use batch sizes of 256 and 16 respectively. We use learning rates of 1e-5 for our RoBERTa encoders and 1e-4 for our linear classifier heads. We train on 4 V100 GPUs up to 20 epochs with early stopping (within 6 epochs).[5]

6.1 Rationale Selection Evaluation

We evaluate predicted rationales on the Check-COVID dev set by considering rationale sentences individually and in aggregate. When considering a rationale's sentences individually (sentence level), we calculate two scores: **standard** precision and recall and the **global recall-focused** (GRF) variant of precision and recall used by the SciFact paper. GRF scores only consider a selected sentence for a claim to be correct if all of the claim's gold sentences are also selected. While we include this score, we show that our full pipeline performs equally well even in cases where all gold sentences are not predicted by the *Rationale Selection* module, suggesting that this focus on recall may be misplaced.

When considering each rationale in aggregate (rationale level), we calculate two different scores: a **strict** score where a predicted rationale is correct only if it is identical to the gold (no additional/missing sentences) and a more permissive **intersection** score where a predicted rationale is correct if it contains at least one gold sentence. As we shall see, predicted rationales that only intersect with the gold rationale still result in good downstream performance on label prediction.

Results In Table 2, we present the baseline performance for the *Rationale Selection* module on Check-COVID's dev set. When evaluating on *composed* claims, models trained on both *composed*

[5] Results are averaged across three random seeds

Label Prediction	**Composed**				**Extracted**			
	REFUTE	NOTENINFO	SUPPORT	Macro F1 (std)	REFUTE	NOTENINFO	SUPPORT	Macro F1
** Train & evaluate on composed **or** extracted separately*								
Check-COVID	85.86	78.91	81.30	82.03 (1.47)	45.01	68.25	70.94	61.40 (14.41)
FEVER + Check-COVID	85.03	80.24	84.85	83.37 (0.93)	84.30	87.03	84.25	**85.19**†(2.11)
SciFact + Check-COVID	82.19	79.00	81.26	80.82 (1.29)	77.33	80.49	82.17	80.00 (2.24
FEVER + SciFact + Check-COVID	85.00	82.05	84.21	**83.76**†(0.63)	82.76	78.44	82.77	81.33 (1.29)
COVID-Fact + Check-COVID	84.58	78.49	85.73	82.93 (0.99)	38.75	34.64	57.73	43.70 (25.17)
** Train on composed **and** extracted together.*								
Check-COVID	83.78	76.91	79.74	80.14 (2.31)	74.59	89.92	84.63	**83.05**†(1.80)
FEVER + Check-COVID	84.29	81.24	83.22	82.92 (1.27)	80.39	79.03	80.69	80.04 (1.38)
SciFact + Check-COVID	83.51	79.49	80.64	81.21 (1.06)	80.03	82.95	83.49	82.16 (1.14)
FEVER + SciFact + Check-COVID	84.17	81.13	84.66	**83.32**†(2.01)	78.94	76.89	81.79	79.21 (2.64)
COVID-Fact + Check-COVID	83.75	77.75	82.86	81.46 (0.99)	72.69	82.49	81.56	78.91 (2.97)
GPT-3.5 (text-davinci-003)	60.22	70.71	75.00	68.64	33.33	54.55	66.67	51.52

Table 3: F1 (avg. of 3 seeds, std in parentheses) of our *Label Prediction* module under different training configurations (combinations of Check-COVID, FEVER, SciFact and COVID-Fact; *composed* and *extracted* claims alone and together).

Single Sentence Rationale Investigation		Predicted Rationale	Sampled Gold Rationale
Allow NEI	Accuracy	85.71	50.00
	Macro F1	58.04	44.85
Ignore NEI	Accuracy	92.86	85.71
	Macro F1	92.51	85.71

Table 4: When a dev set gold rationale contains many sentences but our model only selects one, we compare *Label Prediction* performance using the selected sentence vs. a randomly sampled gold rationale sentence.

and *extracted* claims outperformed models trained on *composed* alone. The model fine-tuned on only Check-COVID and the model fine-tuned on FEVER and Check-COVID performed best on the harder evaluation metrics: 38.31 for *global recall-focused* F1 at the sentence level and 31.13 for *strict* accuracy at the rationale level respectively. However, on the easier metrics, the model fine-tuned on FEVER, SciFact and Check-COVID together scored highest: 52.18 for *standard* F1 at the sentence level and 75.16 for *intersection* accuracy at the rationale level.

Interestingly, we observe that including *composed* claims during fine-tuning degrades performance on *extracted*. Training on FEVER and Check-COVID produces the best scores on *extracted* across both easier and harder metrics.

We note that for both *composed* and *extracted* claims, our best performing models exhibit a 40-point improvement when evaluating with *intersection* accuracy instead of *strict*. It is possible that these high *intersection* scores are masking cases where the *Rationale Selection* module is picking less informative evidence. To understand this better, in Table 4 we show that even when the model only chooses one sentence from a multi-sentence rationale, it is picking high quality evidence. Scores with the model's selections outperform randomly sampled gold evidence by a wide margin.

Finally, our results show that fine-tuning on FEVER in addition to Check-COVID most often boosts performance while SciFact does not always help. This suggests that while the large size of FEVER helps, the genre difference between SciFact claims (from journals) and Check-COVID claims (from news) limits effective transfer.

6.2 Label Prediction Evaluation

We evaluate the *Label Prediction* module on the Check-COVID dev set using standard macro-F1.

Results The baseline results for the *Label Prediction* module are presented in Table 3. For *composed* claims, we note that the model trained on FEVER + SciFact + Check-COVID (*composed* training examples) achieves the best performance (83.76 macro F1). For *extracted* claims, training on FEVER + Check-COVID (*extracted* examples) achieves the highest score (85.19 macro F1). We suspect that FEVER exhibits the best transfer performance due to its large size and its inclusion of the NOTENOUGHINFO veracity class, however, there is still considerable room for improvement, perhaps because the evidence in FEVER is drawn from Wikipedia rather than scientific journals. Additionally, training on both *composed* and *extracted* examples does not improve performance in most settings when compared to training on *composed* or *extracted* claims alone.

			Oracle			Vespa	
Train / Dev	Test	REFUTE	SUPPORT	Macro F1	REFUTE	SUPPORT	Macro F1
* *Allow* NOTENOUGHINFO *predictions*							
Composed	Composed	87.80	86.27	87.04	58.59	68.09	**63.34**
Comp + Extr		83.76	86.27	85.02	54.00	62.22	58.11
Extracted	Extracted	95.65	88.89	**92.27**	40.00	59.26	49.63
Comp + Extr		85.71	87.18	86.45	28.57	59.26	43.92
* *Ignore* NOTENOUGHINFO *predictions*							
Composed	Composed	92.19	90.91	**91.55**	79.70	74.29	**76.99**
Comp + Extr		90.91	88.68	89.79	74.42	69.72	72.07
Extracted	Extracted	88.00	91.89	89.95	56.00	70.27	63.14
Comp + Extr		85.71	92.68	89.20	25.00	73.91	49.46
GPT-3.5 (text-davinci-003)	Composed	63.92	75.18	69.55	74.45	65.35	69.90
	Extracted	40.00	80.85	60.43	62.5	60.00	61.25

Table 5: Full fact-checking pipeline F1 with different configurations of *composed* and *extracted* claims. *Oracle*/*Vespa* indicates abstract retrieval method. *Rationale Selection* and *Label Prediction* uses Tables 2 and 3 †-ed models.

In Table 3, we present the models' performance broken out by label. We note that performance on REFUTE claims is similar to performance on SUPPORT claims on average, demonstrating that REFUTE claims are not easier than SUPPORT claims for our models. As REFUTE claims are manually negated by our annotators, this suggests this process does not introduce spurious signals that our models learn to exploit.

6.3 Pipeline Evaluation

We evaluate the end-to-end performance of the full fact-checking pipeline on each of the Check-COVID *composed* and *extracted* test sets. We do so under two settings: using *Oracle* (i.e., gold) abstracts and using *Vespa* to retrieve abstracts. We use the *Rationale Selection* and *Label Prediction* models that performed best on the Check-COVID dev set (see the †-ed numbers in Tables 2 and 3). For this evaluation setting we do not allow the *Rationale Selection* model to produce empty rationales.

Our labeled NOTENOUGHINFO examples require passing insufficient evidence to the *Label Prediction* module. However, the *Oracle* abstracts always contain sufficient evidence. Additionally, it is difficult to know *a priori* whether *Vespa* is returning abstracts that lack sufficient evidence. This is primarily because, besides the ones we've selected, the CORD-19 dataset likely includes many more abstracts that contain evidence relevant to any given claim. Therefore, when evaluating the full pipeline, we remove examples labeled NOTENOUGHINFO from our test data. Because our label prediction model was trained on all 3 veracity labels, we evaluate it under two conditions: 1) we allow the model to predict NOTENOUGHINFO (*allow NEI*), and 2) we select the output class by only considering the logits for the SUPPORTS and REFUTES indices in the predictor's output (*ignore NEI*). We use the 2-class (SUPPORT/REFUTE) macro F1 scores of the *Label Prediction* module as the final score for each full pipeline variant.

Results We present the full pipeline results in Table 5. First, consider the *Oracle* abstract setting. For *composed* claims, the best performance (87.04 macro F1 for *allow NEI*, 91.55 for *ignore NEI*) is achieved by training on only *composed* examples. Likewise, for *extracted* claims, training on *extracted* claims produces the best performance (92.27 macro F1 for *allow NEI*, 89.95 for *ignore NEI*). These results show that datasets designed for one type of claim do not necessarily transfer perfectly to the other, even within the same domain. When considering the performance of *Vespa* as our *Abstract Retrieval* module, we observe a fairly large drop in F1 (20 - 30 points in most settings). Additionally, the *Vespa*-based full pipelines benefit more from ignoring when the *Label Prediction* module predicts NOTENOUGHINFO. This suggests that *Vespa* may be retrieving less relevant abstracts that result in more predictions of NOTENOUGHINFO. This result demonstrates the importance of including a NOTENOUGHINFO label, as an *Oracle* for retrieving relevant abstracts does not exist in the wild.

Finally, the models for *Rationale Selection* and *Label Prediction* achieve similar performance on both *composed* and *extracted* claims despite the complexity of the *extracted* claims' surface forms. This suggests that the *Vespa* pipelines' degraded performance on *extracted* claims was due to *Vespa* struggling to retrieve evidence for the difficult news claims in the *extracted* subset.

6.4 In-Context Learning with GPT-3.5

Due to the success of in-context learning across a wide range of NLP benchmarks, we additionally evaluate GPT-3.5 (Brown et al., 2020) on Check-COVID using a few-shot setting.[6] As is evident from Tables 2, 3 and 5, GPT-3.5 performs worse than our trained baseline on rationale selection (by 10-20 points), label prediction (by 15-30 points) and in the full pipeline setting (by 6-20 points). It is possible that without task-specific fine-tuning, it struggles to ground the natural language claims in scientific jargon.

7 Conclusion

We present Check-COVID, a new corpus for fact-checking everyday claims about COVID-19 against evidence from scientific journal articles. While our experimental results establish a strong baseline, they also demonstrate the difficulty in transferring learning from existing fact-checking corpora to this new dataset. In future work, we plan to train a unified model to perform rationale selection and label prediction, mitigating error propagation.

Limitations

Due to time and budget constraints, this work remains limited in a number of important ways. The relatively small size of our corpus and its specificity to COVID-19 necessitates the development of systems with richer inductive biases and the ability to effectively transfer knowledge from related corpora like FEVER. Additionally, due to the large size of CORD-19, the database of scientific literature from which we draw the abstracts in Check-COVID, it is difficult to evaluate abstract retrieval components like Vespa with our annotations. There are likely many abstracts in CORD-19 in addition to the ones we've selected that contain evidence relevant to any given claim, precluding both measurements of abstract precision and the evaluation of NOTENOUGHINFO in the full fact-checking pipeline setting. Finally, as the claims in Check-COVID are drawn from western, English language news sources and annotators, they are likely unrepresentative of the full range of COVID-related (mis)information in need of fact-checking online.

Ethics Statement

There are a few ethical considerations to consider if someone were to use our fact-checking system in the real world. Fact-checking results are entirely dependent on the source of truth used. In this work, we use peer-reviewed scientific journal articles, but a malicious actor could easily swap the backend of our system (Vespa) and instead have it search over some source of information that advances that actor's agenda. This type of use would further propaganda and misinformation. Additionally, many fact-checking processes are used to flag and potentially remove content online. Content moderation has many positive effects in promoting healthy online communities, but again, in the wrong hands with the wrong source of truth, this type of system could be used for unjust censorship. Finally, our claims were collected from news sites that lean toward the left or middle on the United States political spectrum[7]. The system could therefore behave differently on claims from more right-leaning news sources in a way that might favor one set of views over another.

Acknowledgements

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0123. The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

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.*

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.

Alberto Barrón-Cedeno, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: A system to unmask propaganda in online news. In

[6] Experimental details can be found in the Appendix C.

[7] https://www.allsides.com/media-bias/media-bias-chart

Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 9847–9848.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*.

Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. 2020. Covid-19-fakes: a twitter (arabic/english) dataset for detecting misleading information on covid-19. In *International Conference on Intelligent Networking and Collaborative Systems*, pages 256–268. Springer.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2020. Misinformation has high perplexity. *arXiv preprint arXiv:2006.04666*.

Nayeon Lee, Yejin Bang, Andrea Madotto, Madian Khabsa, and Pascale Fung. 2021. Towards few-shot fact-checking via perplexity. *arXiv preprint arXiv:2103.09535*.

Manling Li, Revanth Gangi Reddy, Ziqi Wang, Yi-Shyuan Chiang, Tuan Lai, Pengfei Yu, Zixuan Zhang, and Heng Ji. 2022. Covid-19 claim radar: A structured claim extraction and tracking system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 135–144.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nour Mheidly and Jawad Fares. 2020. Leveraging media and health communication strategies to overcome the covid-19 infodemic. *Journal of public health policy*, 41(4):410–420.

Isabelle Mohr, Amelie Wührl, and Roman Klinger. 2022. Covert: A corpus of fact-checked biomedical covid-19 tweets. *arXiv preprint arXiv:2204.12164*.

Preslav I Nakov, Ariel S Schwartz, and Marti Hearst. 2004. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR*, volume 4, pages 81–88. Citeseer.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications.* " O'Reilly Media, Inc.".

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

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

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic. *arXiv preprint arXiv:2106.03794*.

Mourad Sarrouti, Asma Ben Abacha, Yassine M'rabet, 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.

Tal Schuster, Darsh J Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. *arXiv preprint arXiv:1908.05267*.

Dhwanil Shah, Krish Shah, Manan Jagani, Agam Shah, and Bhaskar Chaudhury. 2022. Concord: Covid-19 numerical claims open research dataset. *Available at SSRN 4222185*.

Gautam Kishore Shahi and Durgesh Nandini. 2020. Fakecovid–a multilingual cross-domain fact check news dataset for covid-19. *arXiv preprint arXiv:2006.11343*.

Megha Sundriyal, Ganeshan Malhotra, Md Shad Akhtar, Shubhashis Sengupta, Andrew Fano, and Tanmoy Chakraborty. 2022. Document retrieval and claim verification to mitigate covid-19 misinformation. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 66–74.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Rutvik Vijjali, Prathyush Potluri, Siddharth Kumar, and Sundeep Teki. 2020. Two stage transformer model for covid-19 fake news detection and fact checking. *arXiv preprint arXiv:2011.13253*.

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

Gengyu Wang, Lawrence Chillrud, and Kathleen McKeown. 2021. Evidence based automatic fact-checking for climate change misinformation. In *6th International Workshop on Social Sensing (SocialSens 2021)*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A News Citance Sources

We extracted citances that contain URLs to medical journals from following news websites: The New York Times, The Washington Post, The Atlantic, CNN, NPR, BBC. We note that while these websites are trustworthy and broadly relied upon by fact-checkers, most of them exhibit a bias to the political left which could constrain the breadth of claims we collect.

B Annotation Interface and Instruction Guide

We present the annotation interface in Figure 3. The detailed annotation instruction guide is available at `https://drive.google.com/file/d/1q6diZgcJquxBZMHdViYn34d8DMVEMVV9/view`. The annotation instruction video is available at `https://youtu.be/mEOcouML9oA`.

C Experiment with GPT-3.5

In our studies involving GPT-3.5, we devised prompts for the module and pipeline evaluation. These prompts comprised a random selection of 2 or 3 examples for in-context learning. For the sake of brevity in this paper, we substituted the actual examples in the prompt templates with the placeholder "example abstract/claim/rational sentence". Our work was carried out on the text-davinci-003 variant of GPT-3.5, using the completion API endpoint with hyperparameters, including a temperature setting of 0.7, a cap on the token count at 256, a top-p value fixed at 1, and both frequency and presence penalties set to 0.

Prompt for Rationale Selection in Module Evaluation

List id(s) of sentence(s) in the abstract that support or refute the claim if they exist. If none apply, return empty list.

Abstract:
0: example abstract sentence
... example abstract sentence
7: example abstract sentence
Claim: example claim sentence
Selected id(s): [7]

Abstract:
0: example abstract sentence
... example abstract sentence
6: example abstract sentence
Claim: example claim sentence
Selected id(s): [4, 5]

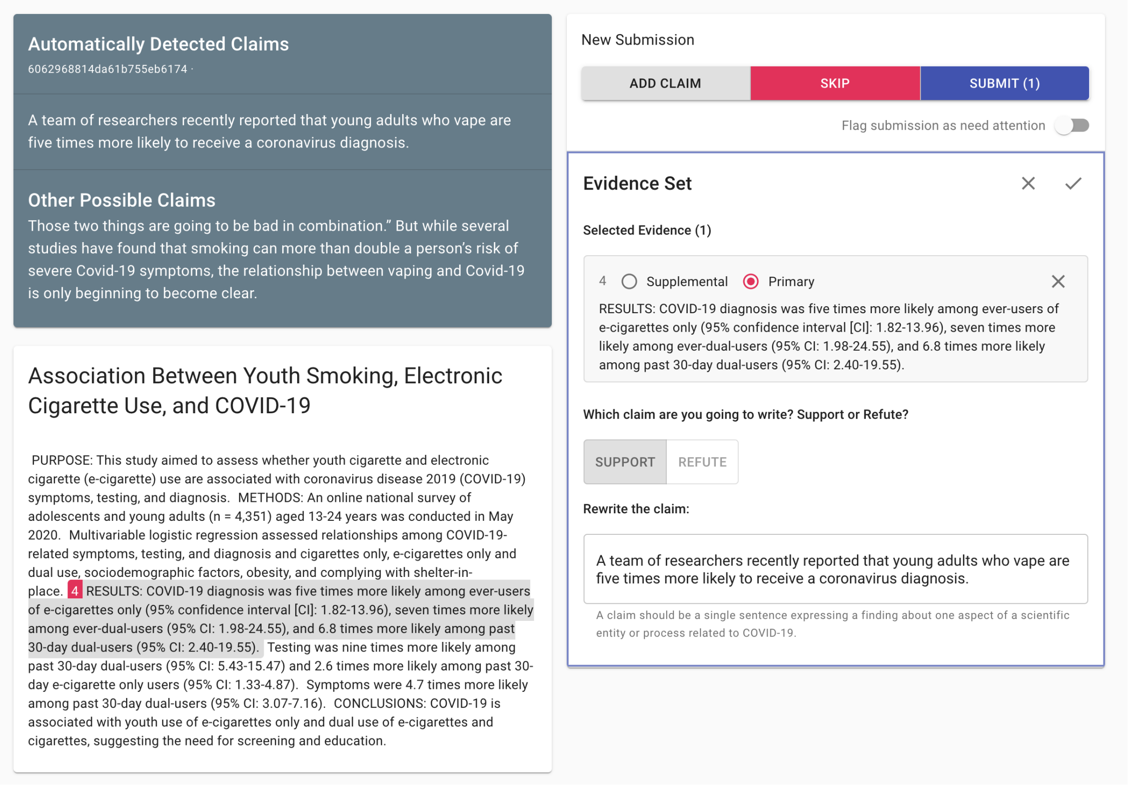
Abstract:
{abstract}
Claim: {claim}
Selected id(s):

Prompt for Rationale Selection in Pipeline Evaluation

List id(s) of sentence(s) in the abstract that support or refute the claim if they exist. If none apply, return at least one that is most related to the claim.

Abstract:
0: example abstract sentence
... example abstract sentence
7: example abstract sentence
Claim: example claim sentence
Selected id(s): [7]

Abstract:
0: example abstract sentence
... example abstract sentence



Automatically Detected Claims

6062968814da61b755eb6174 ·

A team of researchers recently reported that young adults who vape are five times more likely to receive a coronavirus diagnosis.

Other Possible Claims

Those two things are going to be bad in combination.” But while several studies have found that smoking can more than double a person’s risk of severe Covid-19 symptoms, the relationship between vaping and Covid-19 is only beginning to become clear.

Association Between Youth Smoking, Electronic Cigarette Use, and COVID-19

PURPOSE: This study aimed to assess whether youth cigarette and electronic cigarette (e-cigarette) use are associated with coronavirus disease 2019 (COVID-19) symptoms, testing, and diagnosis. METHODS: An online national survey of adolescents and young adults (n = 4,351) aged 13-24 years was conducted in May 2020. Multivariable logistic regression assessed relationships among COVID-19-related symptoms, testing, and diagnosis and cigarettes only, e-cigarettes only and dual use, sociodemographic factors, obesity, and complying with shelter-in-place. 4 RESULTS: COVID-19 diagnosis was five times more likely among ever-users of e-cigarettes only (95% confidence interval [CI]: 1.82-13.96), seven times more likely among ever-dual-users (95% CI: 1.98-24.55), and 6.8 times more likely among past 30-day dual-users (95% CI: 2.40-19.55). Testing was nine times more likely among past 30-day dual-users (95% CI: 5.43-15.47) and 2.6 times more likely among past 30-day e-cigarette only users (95% CI: 1.33-4.87). Symptoms were 4.7 times more likely among past 30-day dual-users (95% CI: 3.07-7.16). CONCLUSIONS: COVID-19 is associated with youth use of e-cigarettes only and dual use of e-cigarettes and cigarettes, suggesting the need for screening and education.

New Submission

Flag submission as need attention

Evidence Set

Selected Evidence (1)

4 ○ Supplemental ● Primary

RESULTS: COVID-19 diagnosis was five times more likely among ever-users of e-cigarettes only (95% confidence interval [CI]: 1.82-13.96), seven times more likely among ever-dual-users (95% CI: 1.98-24.55), and 6.8 times more likely among past 30-day dual-users (95% CI: 2.40-19.55).

Which claim are you going to write? Support or Refute?

[✓ SUPPORT] REFUTE

Rewrite the claim:

A team of researchers recently reported that young adults who vape are five times more likely to receive a coronavirus diagnosis.

A claim should be a single sentence expressing a finding about one aspect of a scientific entity or process related to COVID-19.

Figure 3: Annotation Interface

6: example abstract sentence
Claim: example claim sentence
Selected id(s): [4, 5]

Abstract:
{abstract}
Claim: {claim}
Selected id(s):

Prompt for Label Prediction in Module Evaluation

Decide whether the rationale refutes or supports the claim or if there is not enough info to make a decision on the claim.

Claim: example supported claim sentence
Rationale: example rationale sentences that support the claim
Label: SUPPORT

Claim: example refuted claim sentence
Rationale: example rationale sentences that refute the claim
Label: REFUTE

Claim: example claim sentence
Rationale: example rationale sentences that do not include enough information about the claim to support or refute
Label: NOT ENOUGH INFO

Claim: {claim}
Rationale: {evidence}
Label:

Prompt for Label Prediction in Pipeline Evaluation

Decide whether the rationale refutes or supports the claim.

Claim: example supported claim sentence
Rationale: example rationale sentences that support the claim
Label: SUPPORT

Claim: example refuted claim sentence
Rationale: example rationale sentences that refute the claim
Label: REFUTE

Claim: {claim}
Rationale: {evidence}
Label:

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☑ A2. Did you discuss any potential risks of your work?
Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3 Check-COVID Dataset, Section 5 Baseline System

☑ B1. Did you cite the creators of artifacts you used?
We used an artifact and cite the creators in the first paragraph in Section 5

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Yes, end of Section 1

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 5 and Section 6

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 3.4

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3 and Limitations

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3 and 3.5

C ☑ Did you run computational experiments?

Section 6 Experiments

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 6 Experiments - Hyperparameter Settings

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 6

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 6

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 3.3 and 3.4

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix B Annotation Interface and Instruction Guide

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 3.4

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 3.4

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section 3.4

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 3.4