MSVEC: A Multidomain Testing Dataset for Scientific Claim Verification

Michael Evans
Old Dominion University
mevan028@odu.edu

Dominik Soós
Old Dominion University
dsoos001@odu.edu

Ethan Landers
Old Dominion University
eland007@odu.edu

Jian Wu
Old Dominion University
jwu@cs.odu.edu

ABSTRACT

The increase of disinformation in scientific news across a variety of domains has generated an urgency for a robust and generalizable approach to automated scientific claim verification (SCV). Available methods of SCV are limited in either domain adaptability or scalability. To facilitate building and evaluating more robust models on SCV we propose MSVEC, a multidomain dataset containing 200 pairs of verified scientific news claims with evidence research papers. To understand the capability of large language models on the SCV task, we evaluated GPT-3.5 against MSVEC. While methods of fact-checking exist for specific domains (e.g., political and health), the use of large language models exhibits better generalizability across multiple domains and is potentially compared with state-of-the-art models based on word embeddings. The data and software used and developed for this project are available at https://github.com/lamps-lab/msvec.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Natural language processing**; **Lexical semantics**;

KEYWORDS

benchmark datasets, natural language processing, large language models, machine learning

ACM Reference Format:
Michael Evans, Dominik Soós, Ethan Landers, and Jian Wu. 2023. MSVEC: A Multidomain Testing Dataset for Scientific Claim Verification. In *The Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc '23), October 23–26, 2023, Washington, DC, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3565287.3617630

1 INTRODUCTION

Statistics show that over 50% of Americans consumed news through social media [1]. In combination with the popularity of information retrieval through social media, disinformation shared through this medium spreads six times faster than the rate of true news [1]. With the rise of news consumption on the social Web, it is becoming increasingly difficult to discern factual news from news containing misinformation and disinformation. Automatically detecting and debunking disinformation is a long-standing issue, and various methods have been proposed. Fact-checking websites have helped suppress the spread of disinformation in scientific news, along with the development of automated verification methods utilizing machine learning and natural language processing [10].

The process of verifying scientific news claims with evidence documents is apportioned into two sub-tasks: stance labeling and rationale annotation. Each news claim is generally a one-sentence statement related to a scientific discovery for which an evidence document addresses, such as *Sleep Deprivation Is Surprisingly Effective as an Antidepressant*. Stance labeling refers to determining the stance of a research paper with respect to the scientific news claim (supporting or contradictory). An expert-verified scientific news claim is expected to receive a label of *SUPPORT* by the model, as the accompanying research paper should contain supporting evidence of the discovery. Moreover, a claim encouraging false news should receive a label of *CONTRADICT* by the model, as the related research paper should contain contradictory evidence of a false claim. The second task in SCV, namely rationale annotation, entails labeling individual sentences in the research paper to determine the stance labels (support or contradict). If the model labeled a true claim as being supported by the research paper, sentence rationales indicate the specific sentences used in the decision-making.

Available datasets are typically limited to a focused domain such as health; models trained on these datasets may not be generalizable in unrelated domains. Machine learning models such as MultiVerS [10] and BEVERS [3] build upon pre-trained language models (LMs) to embed claims and evidence documents. These LMs, such as BERT [4] learn natural language patterns and are successful in tasks such as understanding sentiment. However, their performance on SCV tasks across multiple domains has not been widely tested. To support such evaluations, a benchmark dataset covering claims and evidence in multiple domains must be built. Table 1 shows the existing SCV datasets and compares them with the dataset we propose in this paper.

In this paper, we developed a novel benchmark dataset called Multidomain Scientific Claim Verification Evaluation Corpus (MSVEC) containing 200 news-paper pairings used for testing the effectiveness of models in SCV. We outline the process of creating the dataset

MobiHoc '23, October 23–26, 2023, Washington, DC, USA

ACM ISBN 978-1-4503-9926-5/23/10…$15.00
https://doi.org/10.1145/3565287.3617630

and examine the results generated by querying Generative Pretraining Transformer (GPT-3.5) for the tasks of stance labeling and rationale annotation. Our results demonstrate the capabilities of a typical LLM in SCV. Our major contributions are as follows:

(1) Developed a multidomain testing dataset containing scientific claims from news articles with evidence papers and human-annotated rationales. Our dataset contains news claims from 10 domains and consists of 151 true and 49 false claim-paper pairings.
(2) Evaluated the performance of a zero-shot method with GPT-3.5 against the MSVEC dataset on two sub-tasks: stance labeling and identifying sentence rationales.

Table 1: Comparison of MSVEC with existing SCV datasets.

Dataset	# Claims	Domains	Source
SciFact-open	279	Biomedical	Research Papers
HealthVer	230	Covid	Web
Covid-Fact	46	Covid	Web + Generated
MSVEC	200	Multiple	Fact-checking websites + Research Papers

2 RELATED WORK

Several domain-specific datasets exist for training models on SCV. SciFact contained about 1.4K scientific claims and a search corpus of about 5K abstracts that provided either supporting or refuting evidence for each claim [8]. The claims in SciFact-open are extracted from the citation context of research papers in biomedical sciences. SciFact-open includes 279 claims verified against a search corpus of 500K abstracts [9]. SciFact-open takes into account new means of deviation such as varying levels of specificity between the claim and the research abstract.

HealthVer and Covid-Fact are known examples of existing datasets developed on the health domain. HealthVer is a COVID-19 focused dataset that contains 14K evidence-claim pairs manually annotated as support, refute, or neutral. The claims were retrieved from TREC-COVID [7] and from search engines when queried for questions regarding COVID-19; the abstracts were ranked using a T5 relevance-based model [6]. When trained on the HealthVer dataset, the BERT-base model achieved an F1 score of 73.54%. This could be attributed to the fact that the testing data are in the same domain as the training data.

Another domain-specific dataset is COVID-Fact, consisting of 4,086 claims concerning COVID-19 [5]. This dataset was constructed using a unique approach in that all true claims were collected, while all counter-claims were automatically generated by altering the true claims. True scientific news claims were scraped from the r/COVID19 subreddit, which required all posts to include peer-reviewed evidence. From these verified claims, COVID-Fact automatically generated counter-claims to be used as false claims in the training data.

Although the existing datasets for training SCV models perform well on the testing part of the respective dataset, these corpora usually focus on a narrow scope of domains. MSVEC attempts to provide a dataset containing scientific claims across multiple domains of expert-verified news claims with accompanying evidence in the form of research paper abstracts.

3 DATASET CONSTRUCTION

3.1 Data Acquisition

The scientific news claims in MSVEC are scraped from credible scientific news outlets or fact-checking websites, including Snopes.com, ScienceAlert.com, and Reuters.com. This data was obtained by crawling webpages posted from 2014 to 2022, from which posts containing research articles as evidence were hand-selected. After parsing the crawled HTML, only scientific news was selected that contained URLs linking to research articles to back up the justification of the labels. The research articles are identified by URLs containing DOIs or linking to a list of known publishers. All scientific news articles from ScienceAlert were deemed to be true and the references were used as the evidence papers. The titles from ScienceAlert were used as the claims. News articles from Snopes.com were crawled using the sitemap and scientific fact-checking articles backed by research papers were selected as references. Claims extracted from Snopes.com contained an explicit HTML class embedding which was used as the claim of the news article. For Reuters.com, the first paragraph of an article was adopted as the claim. For this pilot study, only news articles labeled as true or false were selected. We identify one evidence paper for each news article.

The resulting dataset consisted of 200 scientific news claims, each linking to a scientific research paper abstract relevant to the claim. This set of claim-abstract pairings served as the ground truth for abstract-level stances of *SUPPORT* or *CONTRADICT*. For the sentence-level rationales, individual sentences of each abstract were manually annotated by a computer science student and served as the ground truth.

3.2 Subject Domains and Metadata

MSVEC includes 10 subject domains as shown in Table 2, namely: Environment, Health, Humans, Nature, Opinion, Physics, Society, Space, Tech, and a small fraction of early news from ScienceAlert was uncategorized. News articles labeled as *mixed* in truthfulness were gathered but were left out of the dataset so only *true* and *false* claims are retained. Metadata for each claim-paper pairing was manually extracted from the source. These metadata fields include news publication date, news and scientific paper URLs, domain, research paper authors, publication date, venue, and title. Domain labels for news claims received from ScienceAlert.com were included in the article, while news claims scraped from Reuters.com required manual annotation by a computer science student for categorizing the domain of the news. Table 2 outlines the properties of the MSVEC dataset.

Table 3 lists the top online libraries by number of research paper contributions. Of the 200 claim-paper pairings, 24.5% of the relevant abstracts were sourced from onlinelibrary.wiley.com. While only 156 of the 200 pairings are accounted for in this table, the additional online libraries each contributed less than 1.5% per web domain, with some libraries contributing a single research paper.

Table 2: Properties of the MSVEC dataset. The number of news-paper pairings is equal to the number of claims.

Subject Domain	# News-Paper Pairs	Percentage	ScienceAlert	Reuters	Snopes/Other	True	False
Health	70	35.0%	9	36	25	47	23
Environment	22	11.0%	7	5	10	13	9
Society	21	10.5%	7	10	4	17	4
Humans	19	9.50%	7	7	5	13	6
Nature	16	8.00%	7	4	5	16	0
Space	12	6.00%	7	1	4	9	3
Tech	12	6.00%	7	2	3	12	0
Opinion	11	5.50%	7	3	1	9	2
Physics	9	4.50%	7	0	2	7	2
Uncategorized	8	4.00%	7	0	1	8	0

Table 3: The top 10 web domains of URLs linking to scientific papers.

Web Domain	# Papers	Percentage
onlinelibrary.wiley.com	49	24.5%
ncbi.nlm.nih.gov	23	11.5%
jamanetwork.com	18	9.00%
sciencedirect.com	18	9.00%
pnas.org	13	6.50%
pubs.acs.org	12	6.00%
tandfonline.com	9	4.50%
bmj.com	6	3.00%
link.springer.com	5	2.50%
science.org	3	1.50%

4 EXPERIMENTS AND RESULTS

4.1 Zero-shot Scientific Claim Verification with GPT 3.5

GPT 3.5 is an LLM with 175 billion parameters, 4096 tokens, and is among the most capable LLMs available [2]. LLMs are generative deep learning models trained with tens of terabytes of data from the Web. We use the gpt-3.5-turbo API to complete both the stance labeling and rationale annotation tasks. For stance labeling, GPT was fed a scientific news claim and research paper abstract pair and was requested to label the stance (*SUPPORT* or *CONTRADICT*) of the paper with respect to the claim. A relevance scoring was also requested (0-1000) based on how relevant the abstract was to the claim but was not used in the confusion matrix of the model. For rationale annotation, GPT was supplied an abstract consisting of numbered sentences and replied with an array of numbers corresponding to the sentences that can be used as stance evidence. During the stance labeling process, the temperature of the model was tested at 0.25, 0.50, and 0.75. Temperature is a hyperparameter controlling the randomness of the generated text, where higher temperatures result in an increased creativity of the responses. Each temperature received three trials and a final consensus was made using a majority voting algorithm applied to the labels. While performing the rationale annotation, each temperature was tested only once.

4.2 Stance Labeling

In this task, GPT was expected to label *SUPPORT* for abstracts relevant to true news claims and *CONTRADICT* for abstracts relevant to false news claims. A third label of *not enough information* (NEI) was allowed as an option for the model, though none existed in the ground truth. Instances in which GPT labeled the abstract as *NEI* were re-queried an additional 3 times with an altered prompt, only allowing for a binary response of *SUPPORT* or *CONTRADICT*.

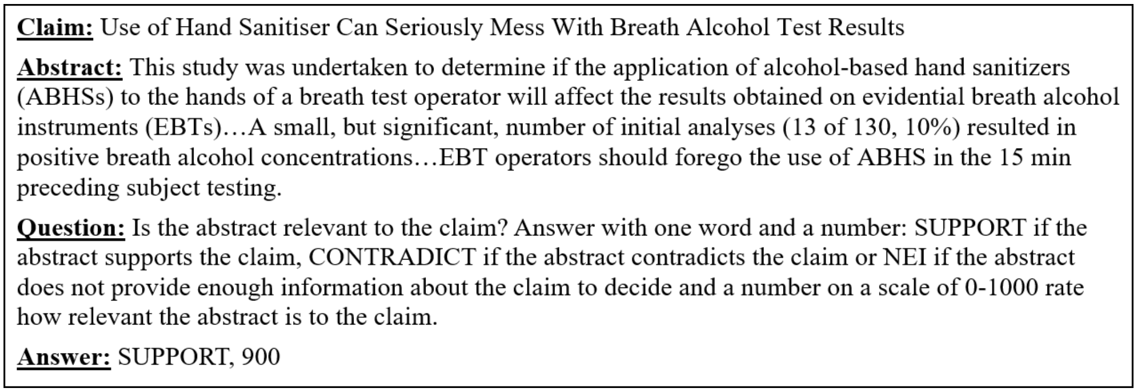
The model was heavily biased towards labeling previous NEI abstracts as *CONTRADICT* after the option for NEI was removed. In 47 re-queries of GPT without the option of *NEI*, 46 responses were labeled as *CONTRADICT* despite a mixture of true and false claims.

The performance of both labels (support and contradict) was measured based on the metrics of precision, recall, and F1 score. Scientific news claims labeled as true in the ground truth and *SUPPORT* by GPT were considered to be true positives (TP) for the support class. False news claims labeled as *CONTRADICT* by GPT were labeled as true negatives (TN). If the label for the claim was false and GPT answered with *SUPPORT*, this was classified as a false positive (FP). True news claims that received a label of *CONTRADICT* by the model were classified as false negatives (FN). The positive label for the confusion matrix was dependent upon which class was being tested; the aforementioned classifications are for the support class specifically. Figure 1 displays an example of the completion of a GPT-3.5 stance labeling prompt and response. This is the original prompt that GPT received. If the *NEI* label was found in the response, we altered the query by removing the *NEI* option.

Results. Table 4 shows the precision, recall, and F1 scores for both classes of the stance labeling task. In both instances, the lowest temperature used in querying GPT (0.25) achieved the highest F1 score for its class. Similarly, both classes performed at their worst when the temperature was at its highest (0.75). However, the model performed substantially worse with the highest F1 being 0.491 when measured by the *CONTRADICT* class, which is 0.144 lower than the *SUPPORT* class.

Consistency. As mentioned above, to mitigate the varied responses from GPT at a non-zero temperature, we submit three queries for each temperature and use a majority voting to decide the final label. Table 5 displays the consistency of the model's responses across each temperature. Consistencies of 1/3 do not exist as claims which

Figure 1: An example stance labeling prompt and response.



Claim: Use of Hand Sanitiser Can Seriously Mess With Breath Alcohol Test Results

Abstract: This study was undertaken to determine if the application of alcohol-based hand sanitizers (ABHSs) to the hands of a breath test operator will affect the results obtained on evidential breath alcohol instruments (EBTs)…A small, but significant, number of initial analyses (13 of 130, 10%) resulted in positive breath alcohol concentrations…EBT operators should forego the use of ABHS in the 15 min preceding subject testing.

Question: Is the abstract relevant to the claim? Answer with one word and a number: SUPPORT if the abstract supports the claim, CONTRADICT if the abstract contradicts the claim or NEI if the abstract does not provide enough information about the claim to decide and a number on a scale of 0-1000 rate how relevant the abstract is to the claim.

Answer: SUPPORT, 900

Table 4: GPT-3.5 stance labeling results. The best performance is highlighted in bold.

Temperature	Class	Precision	Recall	F1
0.25	**SUPPORT**	**0.902**	**0.490**	**0.635**
0.50	SUPPORT	0.900	0.477	0.623
0.75	SUPPORT	0.848	0.444	0.583
0.25	**CONTRADICT**	**0.347**	**0.837**	**0.491**
0.50	CONTRADICT	0.342	0.837	0.485
0.75	CONTRADICT	0.306	0.755	0.435

were initially *NEI* were re-queried with binary label options of *SUPPORT* or *CONTRADICT*. As expected, lower temperatures produced more consistent responses while higher temperatures generated more volatility in the labeling. A majority voting algorithm was used for deciding on a single label from the 2/3 consistency cases. The result indicates that a majority voting on multiple queries is necessary to reduce the randomness of answers given by GPT. Even at a low temperature (0.25), there is nearly a 20% chance that three exact queries do not obtain the same answer.

Table 5: GPT-3.5 consistency of responses for each temperature out of 3 identical queries.

Temperature	Consistency	# Queries	Percentage
0.25	3/3	163	81.5%
0.25	2/3	37	18.5%
0.50	3/3	129	64.5%
0.50	2/3	71	35.5%
0.75	3/3	115	57.5%
0.75	2/3	85	42.5%

Support class results by subject. Table 6 shows the precision, recall, and F1 scores for the support class of the stance labeling task by individual subjects. The *Uncategorized* subject achieved the highest performing F1 score of 0.857, followed by the *Space* and *Tech*. The *Environment* and *Opinion* are tied for the lowest F1 score of 0.500. The value for the Temperature column was chosen based on the best-performing F1 score for each subject.

Table 6: GPT-3.5 *SUPPORT* class stance labeling results by subject.

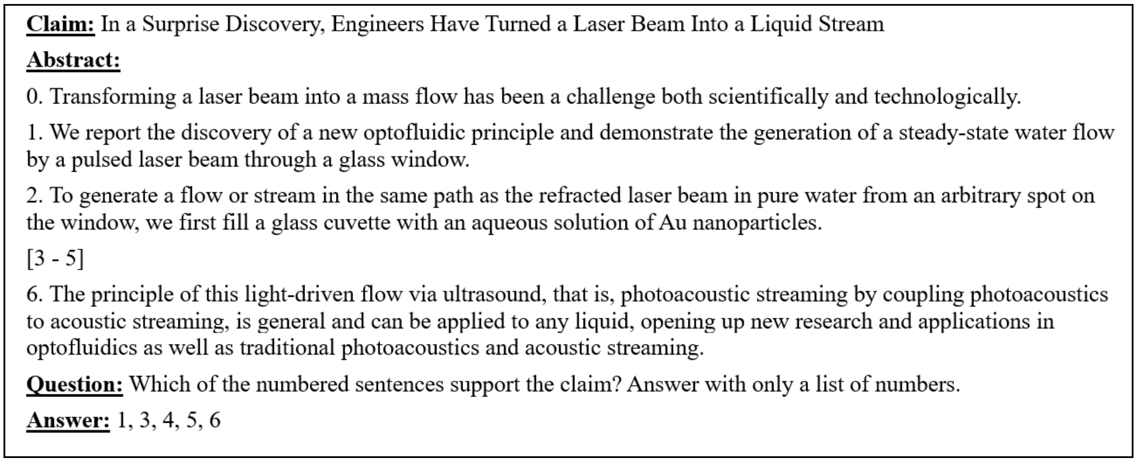
Domain	Size	Temperature	Precision	Recall	F1
Health	70	0.25	0.875	0.447	0.592
Environment	22	0.25 / 0.75	0.714	0.385	0.500
Society	21	0.50	0.917	0.647	0.759
Humans	19	0.75	0.857	0.462	0.600
Nature	16	0.50	1.000	0.563	0.720
Space	12	0.25	1.000	0.667	0.800
Tech	12	0.25	1.000	0.667	0.800
Opinion	11	0.25 / 0.50 / 0.75	1.000	0.333	0.500
Physics	9	0.75	1.000	0.429	0.600
Uncategorized	8	0.25 / 0.50	1.000	0.750	0.857

Contradict class results by subject. Table 7 shows the precision, recall, and F1 scores for the contradict class of the stance labeling task by subject. *Space* achieved the highest performing F1 score of 0.667 while *Nature*, *Tech*, and *Uncategorized* all have undefined F1 scores as these subject domains do not contain false claims (Table 2). The value for the Temperature column was chosen based on the best-performing F1 score for each domain.

4.3 Rationale Annotation

A second prompt was engineered for the task of annotating rationales, along with the formation of an alternative ground truth. Rationale annotation utilized a binary labeling system such that 1 represented an abstract sentence used in determining the stance label, while a 0 indicated that the sentence was not relevant to determining the stance of the abstract with respect to the news claim. A sentence that GPT defined as being a rationale for stance labeling and was labeled as 1 in the ground truth is classified as a TP. If GPT disregarded a sentence (0) that was labeled as a rationale (1), this would be counted as an FN. Similarly, a sentence indicated by GPT as being used in the rationale that was labeled as 0 in the ground truth would be classified as an FP. If GPT disregarded a sentence that was labeled as 0 in the ground truth, this would be

Figure 2: An example of a rationale annotation prompt and response. The notation [3-5] indicates sentences 3 through 5 of the abstract, which are omitted for the purpose of readability in the figure.



Claim: In a Surprise Discovery, Engineers Have Turned a Laser Beam Into a Liquid Stream

Abstract:

0. Transforming a laser beam into a mass flow has been a challenge both scientifically and technologically.

1. We report the discovery of a new optofluidic principle and demonstrate the generation of a steady-state water flow by a pulsed laser beam through a glass window.

2. To generate a flow or stream in the same path as the refracted laser beam in pure water from an arbitrary spot on the window, we first fill a glass cuvette with an aqueous solution of Au nanoparticles.

[3 - 5]

6. The principle of this light-driven flow via ultrasound, that is, photoacoustic streaming by coupling photoacoustics to acoustic streaming, is general and can be applied to any liquid, opening up new research and applications in optofluidics as well as traditional photoacoustics and acoustic streaming.

Question: Which of the numbered sentences support the claim? Answer with only a list of numbers.

Answer: 1, 3, 4, 5, 6

Table 7: **GPT-3.5 *CONTRADICT* class stance labeling results by domain. Domains with empty F1 values because data in such domains do not contain fake news and thus do not have samples in *CONTRADICT* classes.**

Domain	Size	Temperature	Precision	Recall	F1
Health	70	0.25	0.435	0.870	0.580
Environment	22	0.25 / 0.75	0.467	0.778	0.583
Society	21	0.50	0.333	0.750	0.462
Humans	19	0.50	0.429	1.000	0.600
Nature	16	—	0.000	—	—
Space	12	0.25	0.500	1.000	0.667
Tech	12	—	0.000	—	—
Opinion	11	0.25 / 0.50 / 0.75	0.250	1.000	0.400
Physics	9	0.75	0.333	1.000	0.500
Uncategorized	8	—	0.000	—	—

Table 8: **GPT-3.5 sentence rationale results. The best results are highlighted in bold.**

Temperature	Class	Precision	Recall	F1
0.25	Rationale	0.792	0.444	0.569
0.50	Rationale	0.825	0.462	0.593
0.75	**Rationale**	**0.829**	**0.483**	**0.610**
0.25	Non-Rationale	0.282	0.652	0.394
0.50	Non-Rationale	0.306	0.708	0.428
0.75	**Non-Rationale**	**0.313**	**0.704**	**0.433**

counted as TN. Similar to the confusion matrix for stance labeling, the positive class changes while the correctness (true/false) stays consistent. The binary labels used for the ground truth of this task were manually annotated by a computer science student.

Figure 2 is an example of the completion of a GPT-3.5 rationale annotation prompt and response. The abstract of the research paper has been indexed for sentence identification in the model's response.

Rationale annotation results. Table 8 shows the precision, recall, and F1 scores for both labels of the rationale annotation task. As a preliminary study, we only report results to annotate rationales to support true claims. In contrast with the stance labeling task, the highest temperature for both classes achieved the best-performing F1 score. The 0.75 temperature managed an F1 score of 0.610 and 0.433 for the *Rationale* and *Non-Rationale* classes respectively. Table 4 and Table 8 indicate that the temperature does not only affect the randomness of the results but also may affect the average judgment made by GPT. In addition, low temperature does not always make the correct judgment. More experiments are needed to characterize the general dependencies of GPT's responses on SCV queries on temperature.

5 CONCLUSION AND FUTURE WORK

We developed the MSVEC dataset of 200 expert-verified scientific news claims across multiple domains with relevant research papers containing either supporting or contradicting evidence. With this dataset, GPT's performance on stance labeling managed an F1 score of 0.635 in the best-case scenario. GPT tested with MSVEC achieved an F1 score of 0.610 on sentence-level rationales while MultiVerS showed an F1 score of 0.278 during this task when provided zero supervision [10]. Further testing should be done to determine the model's bias towards answering *CONTRADICT* on abstracts that were previously labeled as *NEI*. One limitation of our study was the absence of a human baseline to compare the performance of GPT against. In the future, we propose a human study of students to attempt the task of SCV. Students will perform both tasks of stance labeling and sentence rationales. Human studies will provide an improved understanding of the reliability of GPT.

ACKNOWLEDGMENTS

This project is partially supported by National Science Foundation REU Site Award #2149607 and the Virginia Commonwealth Cyber Initiative Grant #H-4Q21-014.

REFERENCES

[1] David S Ardia, Evan Ringel, Victoria Ekstrand, and Ashley Fox. 2020. Addressing the decline of local news, rise of platforms, and spread of mis-and disinformation online: A summary of current research and policy proposals. *UNC Legal Studies Research Paper* (2020).

[2] 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 (2020), 1877–1901.

[3] Mitchell DeHaven and Stephen Scott. 2023. BEVERS: A General, Simple, and Performant Framework for Automatic Fact Verification. *arXiv preprint arXiv:2303.16974* (2023).

[4] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Vol. 1. 2.

[5] 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* (2021).

[6] 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*. 3499–3512.

[7] Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. TREC-COVID: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, Vol. 54. ACM New York, NY, USA, 1–12.

[8] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974* (2020).

[9] David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-open: Towards open-domain scientific claim verification. *arXiv preprint arXiv:2210.13777* (2022).

[10] David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2021. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. *arXiv preprint arXiv:2112.01640* (2021).

Received 11 August 2023; revised 30 August 2023; accepted 21 August 2023