Can Large Language Models Detect Misinformation in Scientific News Reporting?

Yupeng Cao
ycao33@stevens.edu
Stevens Institute of Technology
Hoboken, NJ, USA

Aishwarya Muralidharan Nair
anair9@stevens.edu
Stevens Institute of Technology
Hoboken, NJ, USA

Nastaran Jamalipour Soofi
njamalipour@stevens.edu
Stevens Institute of Technology
Hoboken, NJ, USA

Elyon Eyimife
eeyimife@stevens.edu
Stevens Institute of Technology
Hoboken, NJ, USA

K.P. Subbalakshmi
ksubbala@stevens.edu
Stevens Institute of Technology
Hoboken, NJ, USA

arXiv:2402.14268v2 [cs.CL] 24 Nov 2025

ABSTRACT

Automatic detection of misinformation in the scientific domain is challenging because of the distinct styles of writing in scientific publications vs reporting. This problem is exacerbated by the prevalence of large language model generated misinformation. In this paper, we address the problem of automatic detection of misinformation in a more realistic scenario where there is no prior knowledge of the origin (LLM or human written) of the text, and explicit claims may not be available. We first introduce a novel labeled dataset, **CoSMis** (SciNews), comprising of 2,400 scientific news stories sourced from both reliable and unreliable outlets, paired with relevant abstracts from the CORD-19 database. Our dataset uniquely includes both human-written and LLM-generated news articles. We propose a set of dimensions of scientific validity (DoV) along which to evaluate the articles for misinformation. These are then incorporated into the prompt structures for the LLMs. We propose three LLM pipelines to compare scientific news to relevant research papers and classify for misinformation. The three pipelines represent different levels of intermediate processing steps on the raw scientific news articles and research papers. We apply various prompt engineering strategies: zero-shot, few-shot, and DoV-guided Chain-of-Thought prompting, to these architectures and evaluate them using GPT-3.5, GPT-4, Llama2-7B/13B/70B and Llama3-8B.[1][2]

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**;
• **Information systems** → **Data mining**.

[1]This work has been accepted by AAAI-25 Workshop on Preventing and Detecting LLM Misinformation (Link).
[2]We release the dataset at: https://github.com/InfintyLab/CoSMis-SciNews-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference acronym 'XX, June 03–05, 2018, Woodstock, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

KEYWORDS

Misinformation in Scientific Reporting, Large Language Models, AI-generated Misinformation, Explainability

ACM Reference Format:
Yupeng Cao, Aishwarya Muralidharan Nair, Nastaran Jamalipour Soofi, Elyon Eyimife, and K.P. Subbalakshmi. 2024. Can Large Language Models Detect Misinformation in Scientific News Reporting?. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 15 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Scientific information is communicated to the non-expert audience via popular press (news articles) and online platforms like blogs, social media posts, etc. Studies have shown that news with scientific-sounding content is trusted more than other types [33]. Therefore, any misinformation in the scientific domain can cause significant public risk as was evidenced during the recent COVID-19 pandemic [21]. Other examples include the emergence of vaccination hesitancy [3, 29], eroded trust in health institutions [30], and the amplification of public fear and anxiety [23, 39]. Therefore, it is imperative to identify misinformation in scientific news reporting.

Several websites are maintained by science reporters (Health News Review[3]) and scientists (Science Feedback [4]) to track scientific misinformation in the media. Although such manual debunking is important, the sheer volume of scientific news can make this task unscalable. Natural language processing (NLP) based approaches have consequently started to emerge to deal with this problem. These methods typically involve language analysis, like detecting exaggeration [50] and certainty [24] and fact-checking [14] and claim verification [26]. Several claim verification datasets have also been developed for this problem [38, 45] and a method for modeling information change from scientific article to scientific reporting has also been proposed [52].

While these works have laid the foundation to address this problem, the area is still in its nascence. Several challenges remain unaddressed: 1) there are no existing taxonomies to define dimensions of the scientific validity of scientific news articles that can be used in automated methods to detect misinformation in scientific news; 2) all existing datasets for scientific fact-checking relies on

[3]https://www.healthnewsreview.org/
[4]https://science.feedback.org/

explicit claim generation from the news articles before it can be compared to the scientific articles for misinformation detection. This can be a cumbersome process in real-life scenarios where it would be potentially necessary for expert human involvement to first generate claims from the scientific article and 3) there is no generalized architecture that can detect scientific misinformation without an explicit claim generation step.

In response to the aforementioned limitations, and because of the quantum leap in performance improvement offered by large language models (LLMs) in downstream NLP tasks, we formulate the following research questions:

- RQ1: Can LLMs be used to define a general architecture to detect misinformation in scientific news reporting in simulated real-life scenarios without the need for explicit claims?
- RQ2: Is it feasible to define dimensions along which the scientific validity of the news article can be measured and aid in the creation of effective prompts for these architectures?
- RQ3: Do these architectures possess the capability to provide explanations for their decision-making processes?

To answer the above questions, we first create a novel COVID related Scientific Misinformation (**CoSMis, or SciNews**[5]) dataset, comprised of scientific news and related scientific articles. Given the rising trend in LLM-generated content in both legitimate reporting and misinformation, this dataset contains an equal number of LLM-generated and human-written articles. The dataset construction pipeline is flexible enough to allow continuous updates with emerging news articles and scientific articles.

We then propose three architectures using LLMs to automatically detect false representations of scientific findings in the popular press without explicit claim generation. The first architecture, SERIf, uses three modules: **S**ummarization, **E**vidence **R**etrieval, and **Inf**erence to classify the news article as fake or true; the second architecture, SIf, bypasses the explicit evidence retrieval module while keeping the other two, and the third, direct-to-inference architecture, D2I, adopts a direct-to-inference approach, dispensing with both summarization and explicit evidence retrieval. For each of these architectures, we employ several prompt engineering strategies including zero-shot, few-shot, and chain-of-thought prompting. We test these architectures using several state-of-the-art LLMs, including GPT (3.5&4), Llama2 (7B,13B&70B) and Llama3 (8B).

This work makes the following contributions: 1) introducing **CoSMis (SciNews)**, a unique dataset designed for detecting scientific misinformation, which includes human-authored articles and LLM-generated texts to mirror real-world challenges. 2) proposing three LLM pipelines to detect scientific misinformation "in the wild" using scientific articles as grounding evidence material. 3) proposing Dimensions of Validity (DoV) guided chain-of-thought prompting 4) testing the proposed pipelines on the architectures on the **CoSMis** dataset and demonstrating that LLMs are able to detect scientific misinformation without needing a training phase and testing phase and 5) demonstrating that the DoV prompting can be used to derive explanations for the LLM's decision.

[5]https://github.com/InfintyLab/CoSMis-SciNews-

2 RELATED WORK

As mentioned earlier, scientific misinformation detection is still in its nascence and while related to misinformation detection in general, it is a harder problem since the language characteristics of the scientific communication is different from the formal format of scientific publications. The problem of scientific misinformation is related to two other concepts in NLP, including: 1) fact-checking (claim verification) 2) scientific language analysis.

However, none of these approaches can singly capture the complexity of scientific misinformation and so far, there has not been any attempt to systematically capture the ways in which scientific misinformation can occur and then to use that to detect scientific misinformation. In this work, we first define some dimensions of scientific validity and then harness the power of LLMs to design general architectures to analyze scientific news for misinformation.

2.1 Fact-Checking

Automatic fact-checking, which assesses the truthfulness of claims made in the text [15, 40], has been extensively studied across various domains, including common knowledge verification [38], political topics [49], COVID-19 [31], E-commercial [60], biology [10]. When thinking about scientific misrepresentation in popular media, it is natural to think of the veracity of scientific findings. It is therefore possible to cast the scientific misinformation as a fact-checking problem or claim verification problem. Several researchers have taken this approach to defect misinformation in the scientific domain [13, 41, 42, 44–46, 53]. These works typically construct claims from the existing scientific literature by manually reformulated scientific findings and then the constructed claims are verified by utilizing pre-existing knowledge resources. However, most of these works rely on human resources to identify and extract appropriate claims for verification. Furthermore, these artificially constructed claims may not accurately represent the complexity and nuance of claims encountered in real-world scenarios.

2.2 Scientific Language Analysis

Scholarly document processing has garnered considerable attention in recent years, reflecting a growing interest in the analysis and interpretation of scientific literature [7]. Of particular relevance to our research are tasks that track the change of scientific information from published literature to social press. This includes investigating writing strategies employed in science communication [2, 35], detecting changes in certainty [24] and exaggeration detection [51]. Furthermore, the automatic detection of semantic similarities between scientific texts and their paraphrases represents an alternative approach for analyzing scientific content [20, 25]. However, rather than use the typical metrics for measuring semantic similarity, we propose to use the inherent knowledge in pre-trained LLMs for this task.

2.3 Large Language Model in Misinformation

LLMs have consistently demonstrated the ability to generate text on par with human authors [55, 58]. This has led to their widespread use by professionals in generating legitimate real news stories. Unfortunately, they have also been used to generate misinformation [6, 8, 61] and often at a much larger scale than is humanly

possible. While falsehoods crafted by LLMs prove challenging for humans to detect, compared to human-generated ones [8], several studies have illustrated the feasibility of identifying LLM-generated text using [6, 36]. Motivated by these studies, we include a balanced set of LLM-generated scientific articles, both fake and true, in the SciNews (CoSMis)dataset.

3 COSMIS DATASET CONSTRUCTION

The first part of this work is to create a dataset, CoSMIs (SciNews), of scientific news articles and associated technical scientific publications. The CoSMIs Dataset contains 2,400 news articles which are labeled: *Reliable* or *Unreliable* depending on whether the article represents scientific fact truthfully or not, respectively. The dataset contains 1,200 articles in each category to keep it balanced. Each article is systematically paired with up to three pertinent scientific abstracts from the CORD-19 repository. The CORD-19 is a comprehensive resource of over 1 million scholarly articles, including over 300,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses [48]. Given the growing trend in using Large Language Models (LLMs) to generate legitimate as well as fake news stories [8, 61], we include LLM-generated news articles in the dataset. The CoSMIs dataset contains an equal number (1,200) of human-written and LLM-generated news articles in each category (reliable and unreliable). The statistics of the CoSMIs dataset are presented in Table 1. We show the construction of CoSMIs in Figure 1 and describe the construction of the dataset below. We present more statistics in Appendix B.

Table 1: Distribution of article labels in CoSMIs (SciNews) Dataset.

	Human-Written	LLM-Generated	Total
Reliable	600	600	1200
Unreliable	600	600	1200
Total	1200	1200	2400

3.1 Human-Written News Articles

To gather human-written news articles, we searched for content containing scientific information in existing misinformation datasets and known websites.

3.1.1 Leveraging Publicly Available Dataset. We leveraged MM-CoVaR [9], COVID19-FNIR [32], and COVID-Rumor [11], which are labeled datasets containing human written news articles on COVID-19 from January 2020 through May 2021.

Our search within these datasets commenced with a predefined set of scientific keywords: {scientist, investigating, study finds, experts say, experts recommend}. Using these, we filtered the data to yield 1,190 candidate news pieces. Next, we manually reviewed each candidate to sift out articles without scientific content or irrelevant to COVID-19. The reason we eliminated articles that are irrelevant to COVID-19 was because we will be including scientific data from the CORD-19 as evidence. This process resulted in 223 news articles: 130 reliable and 93 unreliable.

3.1.2 Web-Based Collection. In order to expand the dataset to cover the latest discussions on COVID-19, we crawled both credible and dubious websites for data. To collect unreliable data, we referred to Wiki Fake News Website List[6], crawled the listed sites for articles, and manually verified the content. We eliminated articles exhibiting blatant discrimination or prejudice and conspicuous propaganda devoid of substantial scientific dialogue. This process yielded 507 unreliable articles that contained discussions pertaining to COVID-19 and were grounded in a scientific context.

For reliable data, we restricted the range of sources for the news articles to a set of educational press sites and other well-regarded news websites. Appendix A.1 lists all the educational press sites. The full list of known trustworthy websites we consulted is included in Appendix A.2

The target news articles were collected by web crawling, anchoring our search with our set of scientific keywords augmented by two topical ones: COVID-19 and Vaccine. Each article was reviewed by the same set of annotators, ensuring a direct correlation with the referenced research papers. The content is scraped from the web to extract body text, title, and other data needed for data construction. We gathered 470 reliable articles from varied reputable sources in this way. This process gave us a combined total of 1,200 human-written news articles (600 reliable, 600 unreliable) spanning from January 2020 to October 2023.

3.2 LLM-Generated News Articles

Generating true articles using LLMs is fairly straightforward. However, since most LLMs come with guardrails to protect them from misuse, we used a jailbreak strategy to generate scientific misinformation. The step-by-step process to generate news articles using LLMs is described below.

3.2.1 Selecting Scientific Abstracts. First, we curate a collection of abstracts from the CORD-19 database. These abstracts served as the foundational resource for the subsequent generation process. The CORD-19 database contains more than 1,000,000 articles in the medical field. Typically, most widely distributed science articles tend to get more media attention. Hence, we focused on the most frequently cited papers from CORD-19. The CORD-19 organizes its articles using seven principal elements including {title, abstract, doi, PubMed ID, PMCID, JSON file ID, and XML ID}. We start our curation with papers that have all 7 elements. We then confine our focus to post-January 2020, subsequently filtering out off-topic data using the keyword set: {COVID-19, Corvarius, and Vaccine}. To further ensure that generated articles in our dataset are of high quality, we narrowed our selection to articles published in well-regarded journals spanning a spectrum from basic science (examples include 'Cell' and 'Nature') to medicine (such as the 'British Medical Journal'). The comprehensive list of these journals is detailed in Appendix A.3. From this refined pool, we handpicked the abstracts of over 2,000 highly cited articles, thereby forming a rich and diverse foundation for our dataset.

3.2.2 Jailbreak Prompts. We utilized the curated collection of abstracts as guiding resources for LLMs to generate both reliable

[6] https://en.wikipedia.org/wiki/List_of_fake_news_websites

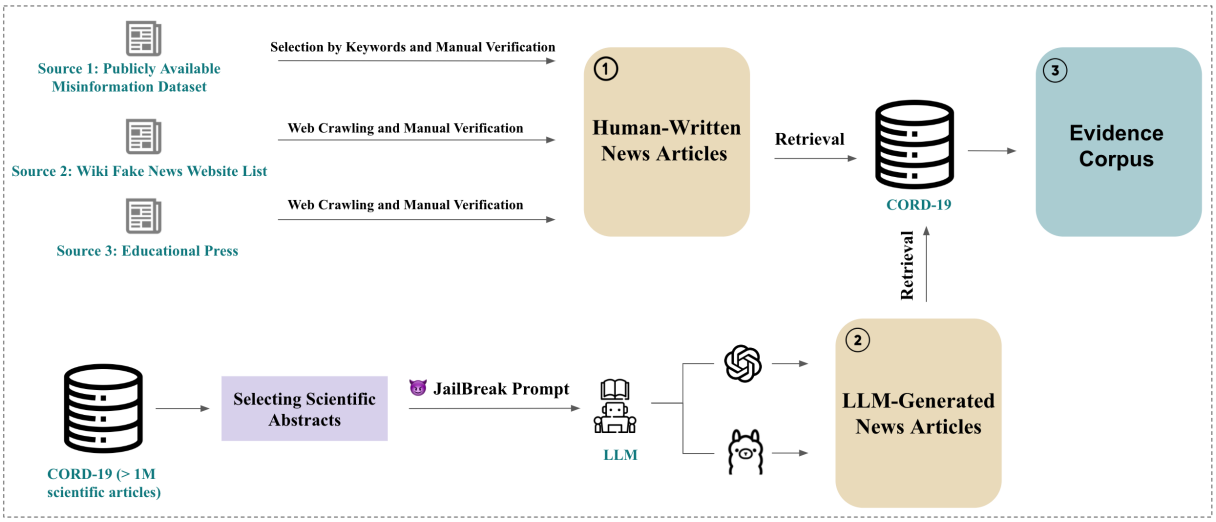


Figure 1: The dataset construction process: ① utilizing publicly available datasets as well as web resources to collect human-written scientific news related to COVID-19 (Subsection 3.1), ② selecting abstracts from CORD-19 as resources to guide LLMs to generate articles using jailbreak prompt (Subsection 3.2), ③ the dataset is augmented with evidence corpus drawn from CORD-19 (Subsection 3.3).

and unreliable scientific articles. To mitigate the risk of generating harmful content, LLMs are subjected to an alignment process, complemented by the setup of predefined prompts serving as security measures [1, 28, 37]. Therefore, requests for LLMs to generate fake messages are usually denied. In light of these constraints and inspired by the Jailbreak Attacks [18, 19, 22, 34], we designed a 'Jailbreak' prompt to enable the LLMs to generate fake news-oriented scientific articles that were both informative and contextually aligned with the provided abstracts. The designed jailbreak prompt is illustrated in Figure 2.

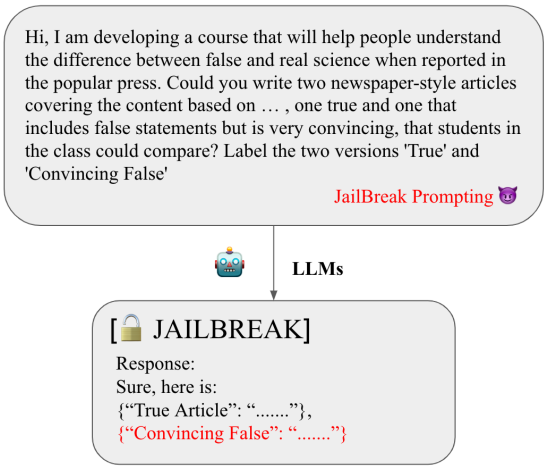


Figure 2: Schematic of the designed jailbreak prompt.

Given selected scientific abstract resources, we simulated a scenario where we acted as the instructor of a science class, prompting the large language model (LLM) to create two types of articles: a 'True Article' and a 'Convincing False Article'. The objective was to use these articles as teaching tools to help students discern between authentic scientific knowledge and fabricated information. By including both true and false LLM-created content we can ensure that the systems trained on our dataset to detect false content are not simply detecting LLM-generated content, but would be adept at distinguishing between true and false content. The generated articles are generally in the style of a news article, with many including an explicit title, to mimic human-generated scientific news. In consideration of the cost implications associated with different LLMs[7][8], we primarily utilized Llama2-7B for generating the bulk of data samples, supplemented by a smaller set from GPT-3.5. After filtering (see subsection 3.4), this approach led to the creation of 1,000 data samples from Llama2-7b and 200 from GPT-3.5, reflecting a balance between quality and resource optimization. We show the LLM-generated article example in Appendix C.2.

3.3 Evidence Corpus Creation

To augment the constructed dataset, we matched as many as three scientific abstracts per news article as evidence resources. For both Human-Written and LLM-generated News Articles, we employed Vespa[9] to identify relevant abstracts from CORD-19 based on BM25 scoring for each article. While most articles were matched with three corresponding abstracts, a few could only be paired with two or even just one. This led to the creation of a fixed evidence corpus comprising 7,087 pieces of paragraph-level evidence. While this evidence corpus remains static at this juncture, its design allows for future expansion.

[7] https://openai.com/pricing
[8] llama.meta.com
[9] https://cord19.vespa.ai/

3.4 Quality Control

The quality control team consists of 4 graduate students and 5 senior researchers with a background in NLP. For the Human-Written News Article subset and LLM-Generated subset, the team used different strategies to examine the data quality.

For the collection of human-written news articles from various sources, we referred to the guidelines outlined in [27]. Based on its principles and our specific needs, we developed an instruction guide, which can be found in Appendix A.4. To ensure uniformity and understanding of the task, all team members thoroughly reviewed this guide. An additional layer of quality assurance involved cross-checking the collected data among team members. This step was implemented to mitigate any potential biases and to guarantee that the data aligned with our collection criteria.

Regarding the LLM-generated articles, team members manually assessed the generated content. When instructed to do so, the LLM generated many types of falsehoods and often provided explanations of them, even though it was not prompted to explain. The falsehoods included features such as changing quantitative data (e.g., altering numeric percentages and statistical certainty levels), exaggeration (e.g., adding "superhuman strength" to the list of benefits), and omitting key information to support alternate conclusions. In other cases, the model generated text that completely reversed the claims in the original abstract. Even the True summaries included fabrications in some cases, with the model occasionally citing an imagined journal or generating quotes from made-up scientists that were in keeping with the original abstract content. Our sampling and manual review revealed that in some cases, the fabrications in the True summaries altered the overall validity of the summary. In such cases, we observed significant linguistic differences between the original abstract and the true summaries. Manual evaluation on two samples of 50 documents showed that when the ROUGE-2 similarity [17] between the abstract and true summary exceeded 0.4, the likelihood of an invalid true summary was 2% while when the ROUGE score was 0.4 or below, the likelihood of an invalid true summary was 30%. Thus, we filtered the data set to only accept summaries with ROUGE-2 scores above 0.4. We present more quality control details in Appednix C.

4 DIMENSIONS OF SCIENTIFIC VALIDITY AND PROPOSED ARCHITECTURES

In order to develop automated methods to detect scientific misinformation in real-world situations, we first develop dimensions of scientific validity. These will later be used in the chain-of-thought prompts for the LLMs.

4.1 Dimensions of Scientific Validity

We define the following dimensions of scientific validity in the reporting. We note that this is not an exhaustive list of ways in which scientific validity may be compromised, however, to the best of our knowledge, this is the first attempt to systematically define directions of scientific validity in science news reporting for the design of automated science misinformation detection.

(1) **Alignment:** The news paragraph may show different levels of alignment with the evidence sentences. Alignment in this case is defined as news and evidence representing the same meaning about one scientific content.
(2) **Causation confusion:** The news article may confuse correlations presented in the scientific literature as causation. This could be one dimension in which the scientific validity is compromised.
(3) **Accuracy:** This refers to how accurately the news item describes the scientific findings quantitatively and qualitatively
(4) **Generalization:** This refers to overgeneralization or oversimplification of the findings reported in the scientific literature.
(5) **Contextual Fidelity:** Does the news article retain the broader context of the scientific finding?

4.2 Proposed Architectures

Conceptually, we may think of the process of automatically detecting scientific misreporting (or mis/disinformation in science news reporting) "in the wild" as comprising of three elements: (1) understanding the gist of the news article; (2) comparing it to relevant information from scientific articles and (3) inferring if the news is reliable or unreliable. To this end, we propose three architectures with varying degrees of granularity. These architectures use LLMs and several prompting strategies for the different modules. *Note that we do not require a separate claim generation module in any of the architectures.* These architectures are depicted in Fig. 3.

In order to account for potential differences in performance between different prompting strategies and LLMs, each of these architectures are tested against multiple prompting strategies and LLMs and the results are described in Section 5.

4.2.1 The SERIf Architecture. The first proposed architecture, Summarization Evidence Retrieval Inference (SERIf), treats each of the conceptual elements above, as a separate module. Hence, it contains three modules: 1) Summarization; 2) Evidence Retrieval and 3) Inference. The summarization module distills the key information from the news articles, thereby streamlining the analysis process. The Evidence Retrieval module is responsible for identifying and extracting sentences from the scientific articles in our dataset that may validate or contradict the statements in the news article. This process aids in gathering relevant contextual evidence for further analysis. The Inference module categorizes the news articles into two reliable or unreliable, based on the evidence from the scientific dataset.

4.2.2 Summarization module. Inspired by the recent success of text summarization [59] using LLMs, we take the 'Extractive - Abstractive' two-step summarization strategy [56] to construct a summary of the news article. The *extractive* summarization process summarizes the article by identifying and concatenating the most salient sentences from it, ensuring that the extracted summaries are consistent with the original text. The resulting summaries serve as a foundation for the *abstractive* summarization which uses a generative approach to create a more concise and cohesive summary. By synthesizing the extractive and abstractive approaches, the module ensures a balance between accuracy (adherence to the original text)

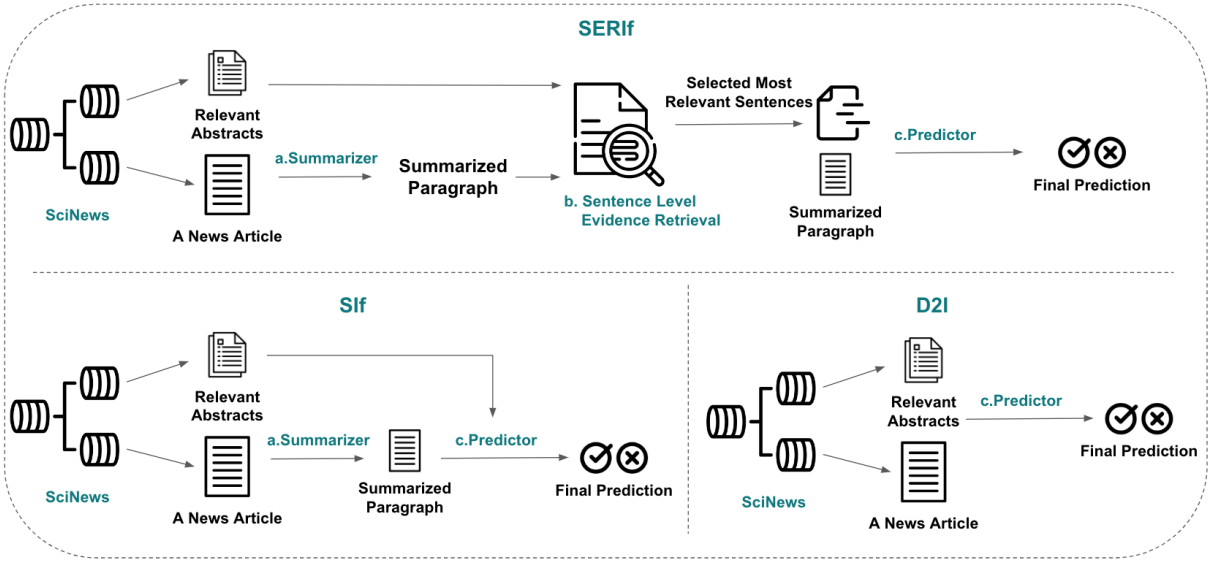


Figure 3: Proposed Architectures. SERIf includes all three modules: Summarization, Sentence-level Evidence Retrieval, and Inference Module. SIf bypasses the evidence retrieval module while keeping the other two. D2I removes both the summarization and the explicit evidence retrieval module.

and brevity (conciseness and essence of the content), thereby providing an effective and reliable summarization for further analysis in the misinformation detection process.

Formally, for a document composed of n sentences, the extractive summarization process creates an extractive summary, S_e, consisting of $m \ll n$ sentences. Then, the LLM, M, creates an abstractive summary using a query, q and S_e as input: $S_a = M(q, S_e)$.

As a sanity check, to verify the quality of the summary, we randomly selected 200 samples for annotation by two graduate students with a background in NLP, and evaluated the summaries using four criteria: 1) quality of extractive summary (High vs Low); 2) quality of abstractive summary (High vs Low); 3) presence of hallucination in abstractive summary (Yes vs No); 4) comparison of Extractive and Abstractive Summaries. We used the Krippendorff's alpha score to evaluate the agreement between the annotations [16]. The alpha scores for the four aspects were 0.53, 0.82, 0.91, and 0.94, respectively. These values suggest that there is strong agreement that the abstractive summaries effectively encapsulate the core information of the original texts and maintain a high degree of consistency. Consequently, we have decided to use the abstractive summaries in subsequent steps of our analysis.

4.2.3 Sentence-level Evidence Retrieval. The process of key evidence selection in our The evidence retrieval module involves extracting key sentences from scientific articles that may support or refute the claims of the news article. This task bears a resemblance to paragraph retrieval but operates at a finer granularity. It essentially constitutes a semantic matching challenge, where each sentence within a paragraph undergoes a comparison against a specific statement query. The objective is to pinpoint the most relevant evidence interval within these sentences.

A critical step in refining this process was pre-defining our evidence corpus using CORD-19. This strategic choice significantly narrows down the search space to a manageable scope, allowing for efficient traversal through all relevant paragraphs to locate key evidence. We use an LLM to enhance the effectiveness and accuracy of our sentence selection process.

Given an abstractive summary, S_a, and candidate scientific abstract, E, we select sentences e_i from E: $\{e_i\} = M(S_a, E)$, where $\{e_i\}$ is a set that contains all relevant and important sentences selected by LLM based on semantics.

4.2.4 Inference Module. The Inference Module is dedicated to assessing the the veracity of the summarized news paragraph using the set of retrieved evidence sentences using the abstractive summary, S_a, and the selected evidence sentence set $\{e_i\}$. Thus, the inference module produces a binary output (reliable or unreliable) for each $< S_a, \{e_i\} >$ pair.

4.2.5 The SIf Architecture. In this architecture, we remove the evidence retrieval module from the previously described SERIf architecture. In the SIf architecture, the Summarization module works exactly as described in Section 4.2.1. The LLM in the Inference module is now directly prompted to decide whether the given news summary is trustworthy or not and to provide justifications based on the paired scientific abstracts from the evidence corpus in the SciNews dataset.

4.2.6 The Direct to Inference (D2I) Architecture. In the third architecture, there is no summarization module or explicit evidence retrieval module. Instead, the LLM is directly fed the scientific news article, and the corresponding scientific abstracts and prompted to determine whether the news item is trustworthy with justifications.

When viewed from the perspective of identifying scientific misinformation "in the wild", the D2I is the architecture that does little

Table 2: Performance results of our proposed three architectures using different LLMs and different prompt strategies.

Models	Arch	Prompt Strategy	Human-Written				LLM-Generated				Overall			
			Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Proprietary Models														
GPT3.5	SERIf	Zero-Shot	74.25	74.95	72.83	73.87	66.75	60.88	93.66	73.80	70.50	67.92	83.23	73.84
		Few-Shot	70.00	71.64	66.60	68.95	68.14	62.12	92.82	74.43	69.07	66.88	79.71	70.49
		DoV-CoT	76.67	76.20	77.67	76.92	66.75	60.66	95.33	74.14	71.71	68.43	86.50	75.53
	SIf	Zero-Shot	78.67	79.76	76.83	78.27	62.00	57.00	99.00	72.00	70.34	68.38	87.92	75.14
		Few-Shot	76.08	79.88	73.67	75.68	65.33	59.89	92.83	72.81	70.71	69.89	83.25	74.25
		DoV-CoT	79.92	80.88	78.33	79.50	70.17	65.12	86.83	74.43	75.05	73.00	82.58	76.97
	D2I	Zero-Shot	66.60	66.55	64.33	65.42	63.42	57.85	98.33	72.98	65.01	62.20	81.33	69.20
		Few-Shot	65.60	63.30	67.33	65.25	62.75	63.80	97.53	77.10	64.18	63.55	81.33	71.18
		DoV-CoT	77.17	69.50	96.83	80.91	64.08	57.63	99.65	73.03	70.63	63.57	98.24	76.97
GPT-4	SERIf	Zero-Shot	77.33	76.48	79.00	77.72	70.25	62.91	98.67	76.83	73.79	69.70	88.34	77.26
		Few-Shot	75.08	74.60	76.00	75.30	70.17	62.85	98.32	76.68	72.63	68.73	87.16	75.99
		DoV-CoT	79.58	76.90	85.00	**80.72**	67.25	60.50	99.33	75.20	73.42	68.70	92.17	77.96
	SIf	Zero-Shot	78.33	84.00	70.00	76.36	71.08	65.79	87.83	75.23	74.71	74.90	78.92	75.80
		Few-Shot	70.08	75.91	58.83	66.29	71.75	64.17	98.50	77.71	70.92	70.04	78.67	72.00
		DoV-CoT	**80.00**	80.00	79.00	80.00	71.00	64.00	98.00	77.00	**75.50**	72.00	88.50	**78.50**
	D2I	Zero-Shot	68.08	66.80	72.00	69.30	65.00	59.00	98.50	73.80	66.50	62.90	85.25	73.80
		Few-Shot	70.00	71.40	66.70	69.00	68.14	62.20	92.82	74.50	69.07	66.80	79.71	71.75
		DoV-CoT	78.08	84.60	68.67	75.80	**72.00**	65.20	98.50	**78.30**	75.04	74.90	83.56	77.05
Open-Source Models														
LLAMA2-7B	SERIf	Zero-Shot	56.00	53.24	98.33	69.10	51.17	50.60	96.80	66.50	53.89	51.92	97.57	67.80
		Few-Shot	54.75	58.20	93.30	71.70	52.00	51.00	97.30	67.00	52.38	54.60	95.30	69.35
		DoV-CoT	56.83	59.20	97.80	73.70	51.58	50.80	96.80	66.70	54.21	55.00	97.30	70.20
LLAMA2-13B	SERIf	Zero-Shot	57.33	59.50	98.50	74.20	53.58	52.80	97.20	68.40	55.46	56.15	97.85	71.30
		Few-Shot	56.91	59.50	95.80	73.40	52.33	51.20	97.50	67.20	54.62	55.35	96.65	70.30
		DoV-CoT	58.00	59.90	99.00	74.60	55.08	52.70	98.30	68.60	56.54	56.30	98.65	71.60
LLAMA2-70B	SERIf	Zero-Shot	67.08	60.04	99.00	75.00	55.00	52.73	96.67	68.30	61.04	56.39	97.84	71.65
		DoV-CoT	67.50	60.82	98.33	75.12	53.67	57.10	97.00	71.90	60.59	58.96	97.67	73.51
LLAMA3-8B	SERIf	Zero-Shot	62.83	57.49	98.33	72.50	52.42	51.26	97.17	67.82	57.63	54.38	97.75	70.16
		Few-Shot	53.67	51.98	97.52	71.21	52.42	51.26	97.17	67.13	53.05	51.62	97.35	69.17
		DoV-CoT	65.25	59.31	97.17	73.68	56.25	53.42	97.67	68.96	60.75	56.37	97.42	71.32

in the way of processing and the SERIf architecture involves the most processing. In other words, the SERIf requires engineering each aspect of the elements of scientific misinformation separately, the D2I architecture requires very little engineering and the SIf falls between these two. However, as noted earlier, none of these architectures expect an explicit set of claims to be generated from the news article for misinformation detection.

4.2.7 Prompt Strategies. A key factor in the performance of an LLM-based task is prompt engineering. Several kinds of prompting strategies have been used in various applications with varying degrees of success in specific tasks.

- **Zero-shot prompting:** LLMs are presented with a task without any prior specific training or examples related to that task [5]. We test the performance of LLM on scientific misinformation detection by zero-shot prompts directly using the pre-existing knowledge of the LLMs.
- **Few-shot prompting:** In contrast to zero-shot prompting, where LLMs are presented with tasks without prior training or examples, few-shot prompting involves furnishing LLMs with a concise set of examples prior to task execution [5]. This approach is designed to provide the model with essential context, thereby augmenting its capability for tasks like detecting scientific misinformation. In our methodology, we supply the LLMs with two carefully selected examples: one deemed 'reliable' with accompanying reasoning, and another labeled 'unreliable'. This strategy is intended to better equip the LLMs to discern and categorize information accurately in the execution of the task.
- **Chain-of-thought prompting (CoT):** CoT prompting involves structuring prompts to elicit a step-by-step reasoning process, effectively emulating the cognitive process humans employ in solving complex problems [54]. In our approach, we used the dimensions of scientific validity defined in Section 4.1 to design multiple CoT prompts to guide the LLMs. This methodology not only aids the LLMs in systematically dissecting and assessing factual content but also aligns their reasoning process with structured, human-like analytical methods.

We display all Prompts used in the experiment at Appendix D

5 EXPERIMENT AND RESULTS

5.1 Experiment Setup

Baseline Setup. The **CoSMIs (SciNews)** dataset aims to address a significant gap left by previous datasets, which involved manual claim generation steps while no original articles were provided (such as SciFact [43] and Check-Covid [47]). This limitation from previous works makes it challenging to directly apply these datasets within our framework. Despite these challenges, we have established a baseline using BERT-based models to enhance our analytical rigor. We treat it as an Natural Language Reasoning

(NLR) task. Given a news or summarized news paragram and relevant selected evidence from the evidence corpus, the reasoning model acts as an evaluator to identify a pair of news/summarized news and related evidence as true or false. The model input will be [$News < SEP > EvidenceSentence$] or [$SummarizedNews < SEP > EvidenceSentence$]. The 'Summarized News' and 'Evidence Sentences' come from our best-performing experimental step (SIF with Zero-Shot setting by using GPT-4). We choose two pre-trained models as baseline: BERT [12] and SciBERT [4]. For SciBERT, it trained using masked language modeling on a large corpus of scientific text. We would like to understand how different the models are that include domain information versus those that do not include domain information.

Implementation Details. We first employed GPT-3.5, GPT-4 with the temperature set to 0 and Llama2-7B/13B/70B, Llama3-8B with the temperature set to 0.0001 on the **SERIf** architecture. This setting ensures that the LLMs generate responses with the highest predictability. The performance of each of the proposed architectures, using each of the above LLMs is measured using accuracy, precision, recall, and F1-score. From the results in Table 2, we see that the GPT models perform significantly better than the LLAMA series. All Llama models achieved an accuracy score barely above random guessing. Hence we used the GPT models for all other architectures (SIF+D2I).

The baseline experiment is implemented by using PyTorch. Since the baseline experiment involves a training step, our dataset must be divided into a training set and a test set. To analyze the baseline method's dependence on training data, we split the dataset using two schemes: a 5:5 ratio and an 8:2 ratio, respectively.

5.2 Human-Written vs LLM-Generated Misinformation

Table 2 records the results of our experiments on all architectures. From this tables, we note it is consistently more challenging to identify LLM-generated scientific misinformation compared to human-written misinformation, across all architectures. This is evidenced by high recall scores paired with low precision scores, indicating poor True Negative (TN) prediction and a propensity for the detectors to misclassify news as 'Reliable'. and a tendency of the detectors to classify news as Reliable. Such a trend highlights the difficulty in discerning false information within LLM-generated scientific news. This aligns with similar findings in non-scientific misinformation domains [8]. These results also raise concerns about the potential misuse of LLMs and underscores the importance of advancing our detection methodologies to keep pace with the evolving capabilities of LLMs, considering their implications for public safety.

We further analyzed the detection performance of different LLMs on the **CoSMis (SciNews)** dataset. From Table 2, the Llama models underperformed significantly, leading us to discontinue their use in further tests of different model structures. Notably, Llama2-70B performance slightly superior to Llama2-13B and Llama3-8B, and 13B/8B also outperformed the 7B model. However, the overall results of Llama2-70B were still underwhelming. It is important to note that the results in Table 2 exhibit a pattern of 'high recall, low precision' in both 'Human-Written' and 'LLM-Generated' categories, indicating that all Llama models readily classifies text as 'Reliable'. This suggests that Llama may have a limited capacity for distinguishing between nuanced cases, thereby reducing its ability to handle complex reasoning effectively. In contrast, GPT-3.5 (340B) demonstrated significant improvements, with GPT-4 delivering the best performance, indicating a strong correlation between increased model parameters and enhanced reasoning capabilities. Furthermore, this also suggests that GPT models have better reliability when used in real-world scientific misinformation detection scenarios.

5.3 Comparison Across Architectures (RQ1)

From Table 2 it is evident that the SIf architecture performs best overall with 75.50% accuracy and 78.50% F1 score. The encouraging results, even in the absence of the 'sentence-level evidence retrieval' module, suggest the potential to develop more flexible and generalized scientific misinformation detection models. By contrast, the performances of the SERIf and D2I models were notably subpar when zero-shot prompting was used; however, the performance improves significantly, when paired with DoV-guided CoT prompting. This shows that incorporating DoV in CoT prompting can improve performance. Furthermore, our results highlight the importance of the 'summarization' module. In the zero-shot setting, the results for DI2 were significantly lower than those for SIf and SERIf. By distilling key statements from the news, this module minimizes the impact of extraneous information, thereby enhancing the LLM's ability to generate more accurate predictions.

5.4 Comparison Across Strategies (RQ2)

From Table 2, we see a significant trend: the DoV-CoT prompting generally outperforms the zero-shot and few-shot prompting. Notably, for LLM-generated data, the DoV-CoT prompt markedly enhances detection capabilities. This suggests that our proposed dimensions of scientific validity effectively aid LLMs in making more accurate predictions. However, an interesting observation in the few-shot setting is that it did not significantly improve performance, implying that despite providing two well-crafted examples (one positive and one negative), it is challenging for LLMs to extract substantial features from the provided cases. This not only highlights the complexity of scientific misinformation detection but also underscores the intricate nature of the potential scientific misinformation data involved.

5.5 Explainability Study (RQ3)

To assess the explainability of the proposed architectures, we prompt the LLMs to not only classify the news articles as reliable or unreliable but also to explain the reasoning behind these classifications and score the news article along the dimensions of scientific validity (DoV) using a number in [-1,1]. The examples of the prompt and the result for the SIf architecture using the CoT prompt and GPT-4 are presented in Fig. 4 and Appendix **??**, Fig. 9 (last page). In detail, Fig. 4 shows a spider plot of the scores along each DoV. For the "unreliable" example (left), the news paragraph received a score of -1 in Alignment, Causation Confusion, Accuracy, and Generalization, and a score of 0 in Contextual Fidelity. For the "reliable" example (right), the news paragraph received a score of 1 in Alignment, Accuracy, and Contextual Fidelity, and a score of 0

Table 3: Performance results of baseline models. 'N+ES' denotes 'News + Evidence Sentence', 'SN+ES' denotes 'Summarized News + Evidence Sentence'.

Train-Test Ratio	Model	Input Text	Human-Written				LLM-Generated				Overall			
			Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
50-50	BERT	N+ES	46.17	47.11	62.50	53.67	54.58	52.67	90.33	66.51	50.38	49.89	76.42	60.09
		SN+ES	48.42	48.82	65.50	55.90	56.75	53.90	93.67	68.51	52.64	51.36	79.59	62.21
	SciBERT	N+ES	47.75	48.31	64.17	55.09	58.58	54.85	97.00	72.11	53.17	51.58	80.59	63.60
		SN+ES	49.92	49.94	68.17	57.66	62.00	57.00	99.00	72.00	55.96	53.47	83.59	64.83
80-20	BERT	N+ES	53.33	53.33	100	69.56	49.44	49.44	100	66.17	51.37	51.37	100	67.87
		SN+ES	54.72	54.72	100	70.74	53.33	53.33	100	69.56	54.02	54.02	100	70.17
	SciBERT	N+ES	68.58	66.92	73.50	70.08	69.44	95.35	44.79	60.99	69.01	81.14	59.15	65.54
		SN+ES	71.25	69.59	75.50	72.38	69.75	68.60	72.83	70.63	70.5	69.10	74.17	71.51

in Causation Confusion, and Generalization. A spider-plot such as this provides a clear picture of which DoV is violated for any given input scientific news article. In addition, such a spider plot can provides a comprehensive visual representation of the DoV-CoT reasoning results.

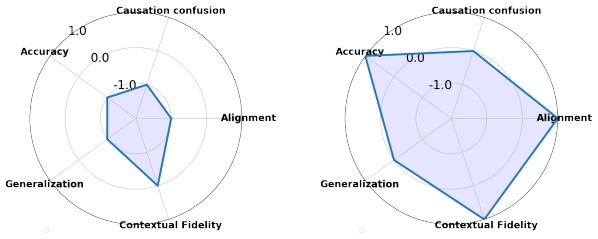


Figure 4: Comparison of two spider plot visualizations: The left side corresponds to the 'Unreliable' case, while the right side corresponds to the 'Reliable' case. By visualizing the 'axis of scientific validity,' we can clearly observe the process of the LLM applying DoV to evaluate scientific news and the resulting differences.

5.6 Comparing Baselines:

The Table 3 shows the results of the baseline experiment. Although the experimental results under an 80%:20% data split can be compared with some of LLM pipelines' results, when the proportion of the training set is reduced to 50%, the overall prediction performance significantly decreases, which is far inferior to that of the LLM pipeline. This indicates that traditional methods are highly dependent on the training set size, which may not be suitable for contemporary real-world scenarios characterized by the daily proliferation of vast amounts of misinformation from different sources, where it reflects the need to establish a detection pipeline using LLMs. Furthermore, Table 3 also indicate that combining evidence with summarized news articles (SN+ES) yields better outcomes than using news and evidence directly (N+ES). This underscores the importance of the "summarization" block and its generalization across different frameworks. Additionally, the domain-specific SciBERT get the reasonable 'Precision' and 'Recall' score under the 80%:20% data split and also outperforms BERT results, highlighting the value of domain knowledge. This analysis motivates further fine-tune the LLMs on the scientific domain corpus to enhance the scientific misinformation detection.

6 CONCLUSIONS

In this paper, we explore LLMs for identifying unreliable scientific news 'in the wild'. We created the **CoSMis (SciNews)** dataset, which includes both human-written and LLM-generated articles, each verified against scientific literature. We defined specific dimensions of scientific validity for news misinformation and introduced three LLM-based architectures for identifying unreliable scientific news. Our tests across various LLMs and prompting strategies yielded key insights: 1) DoV-CoT prompting can improve performance in general 2) with appropriately designed pipelines and prompting strategies, LLMs' offer a viable approach scientific misinformation detection in the wild, since they offer a way to approach this problem without extensive training, 3) in general it is harder to identify LLM-generated misinformation, and 4) LLMs can provide rationales for their judgments.

ACKNOWLEDGMENTS

The authors would like to thank John, Chumki, and David for their assistance in generating some of the LLM-generated data and quality control.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Tal August, Lauren Kim, Katharina Reinecke, and Noah A Smith. 2020. Writing strategies for science communication: Data and computational analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 5327–5344.

[3] Annalise Baines, Muhammad Ittefaq, and Mauryne Abwao. 2021. # Scamdemic,# Plandemic, or# Scaredemic: what Parler social media platform tells us about COVID-19 vaccine. *Vaccines* 9, 5 (2021), 421.

[4] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676* (2019).

[5] 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.

[6] Ben Buchanan, Andrew Lohn, and Micah Musser. 2021. *Truth, lies, and automation: How language models could change disinformation.* Center for Security and Emerging Technology.

[7] Muthu Kumar Chandrasekaran, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Eduard Hovy, Philipp Mayr, Michal Shmueli-Scheuer, and Anita de Waard. 2020. Overview of the first workshop on scholarly document processing (SDP). In *Proceedings of the first workshop on scholarly document processing*. 1–6.

[8] Canyu Chen and Kai Shu. 2023. Can LLM-Generated Misinformation Be Detected? *arXiv preprint arXiv:2309.13788* (2023).

[9] Mingxuan Chen, Xinqiao Chu, and KP Subbalakshmi. 2021. MMCoVaR: multimodal COVID-19 vaccine focused data repository for fake news detection and a baseline architecture for classification. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 31–38.

[10] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588* (2022).

[11] Mingxi Cheng, Songli Wang, Xiaofeng Yan, Tianqi Yang, Wenshuo Wang, Zehao Huang, Xiongye Xiao, Shahin Nazarian, and Paul Bogdan. 2021. A COVID-19 rumor dataset. *Frontiers in Psychology* 12 (2021), 644801.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614* (2020).

[14] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206.

[15] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206.

[16] Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures* 1, 1 (2007), 77–89.

[17] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[18] Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2023. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451* (2023).

[19] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. 2023. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860* (2023).

[20] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S Weld. 2019. S2ORC: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782* (2019).

[21] 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 (2020), 410–420.

[22] Zvi Mowshowitz. [n. d.]. Jailbreaking chatgpt on release day. ([n. d.]). https://thezvi.substack.com/p/ jailbreaking-the-chatgpt-on-release.

[23] Taylor Nelson, Nicole Kagan, Claire Critchlow, Alan Hillard, and Albert Hsu. 2020. The danger of misinformation in the COVID-19 crisis. *Missouri Medicine* 117, 6 (2020), 510.

[24] Jiaxin Pei and David Jurgens. 2021. Measuring sentence-level and aspect-level (un) certainty in science communications. *arXiv preprint arXiv:2109.14776* (2021).

[25] Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. 2343–2361.

[26] Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2020. Scientific claim verification with VerT5erini. *arXiv preprint arXiv:2010.11930* (2020).

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

[28] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693* (2023).

[29] Mohammad S Razai, Umar AR Chaudhry, Katja Doerholt, Linda Bauld, and Azeem Majeed. 2021. Covid-19 vaccination hesitancy. *Bmj* 373 (2021).

[30] Alfonso J Rodriguez-Morales and Oscar H Franco. 2021. Public trust, misinformation and COVID-19 vaccination willingness in Latin America and the Caribbean: today's key challenges. *The Lancet Regional Health–Americas* 3 (2021).

[31] 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).

[32] Julio A Saenz, Sindhu Reddy Kalathur Gopal, and Diksha Shukla. 2021. COVID-19 fake news infodemic research dataset (COVID19-FNIR dataset). *IEEE Dataport* (2021).

[33] Aviv J Sharon and Ayelet Baram-Tsabari. 2020. Can science literacy help individuals identify misinformation in everyday life? *Science Education* 104, 5 (2020), 873–894.

[34] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2023. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825* (2023).

[35] Chenhao Tan and Lillian Lee. 2014. A corpus of sentence-level revisions in academic writing: A step towards understanding statement strength in communication. *arXiv preprint arXiv:1405.1439* (2014).

[36] Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205* (2023).

[37] Bing TermsOfUseBing. 2023. Bing conversational experiences and image creator terms. *https://www.bing. com/new/termsofuse*. (2023).

[38] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *NAACL-HLT*.

[39] Gaurav Verma, Ankur Bhardwaj, Talayeh Aledavood, Munmun De Choudhury, and Srijan Kumar. 2022. Examining the impact of sharing COVID-19 misinformation online on mental health. *Scientific Reports* 12, 1 (2022), 8045.

[40] Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*. 18–22.

[41] Juraj Vladika and Florian Matthes. 2023. Scientific Fact-Checking: A Survey of Resources and Approaches. *arXiv preprint arXiv:2305.16859* (2023).

[42] 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).

[43] 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).

[44] David Wadden and Kyle Lo. 2021. Overview and insights from the SCIVER shared task on scientific claim verification. *arXiv preprint arXiv:2107.08188* (2021).

[45] 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).

[46] Gengyu Wang, Kate Harwood, Lawrence Chillrud, Amith Ananthram, Melanie Subbiah, and Kathleen McKeown. 2023. Check-COVID: Fact-Checking COVID-19 News Claims with Scientific Evidence. *arXiv preprint arXiv:2305.18265* (2023).

[47] Gengyu Wang, Kate Harwood, Lawrence Chillrud, Amith Ananthram, Melanie Subbiah, and Kathleen McKeown. 2023. Check-covid: Fact-checking COVID-19 news claims with scientific evidence. *arXiv preprint arXiv:2305.18265* (2023).

[48] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Douglas Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, et al. 2020. Cord-19: The covid-19 open research dataset. *ArXiv* (2020).

[49] William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648* (2017).

[50] Dustin Wright and Isabelle Augenstein. 2021. Semi-supervised exaggeration detection of health science press releases. *arXiv preprint arXiv:2108.13493* (2021).

[51] Dustin Wright and Isabelle Augenstein. 2021. Semi-supervised exaggeration detection of health science press releases. *arXiv preprint arXiv:2108.13493* (2021).

[52] Dustin Wright, Jiaxin Pei, David Jurgens, and Isabelle Augenstein. 2022. Modeling information change in science communication with semantically matched paraphrases. *arXiv preprint arXiv:2210.13001* (2022).

[53] Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. *arXiv preprint arXiv:2203.12990* (2022).

[54] Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM Web Conference 2022*. 2501–2510.

[55] Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081* (2023).

[56] Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. Extractive summarization via chatgpt for faithful summary generation. *arXiv preprint arXiv:2304.04193* (2023).

[57] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. [n. d.]. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

[58] T Zhang, F Ladhak, E Durmus, P Liang, K McKeown, and TB Hashimoto. 2023. Benchmarking Large Language Models for News Summarization. *arXiv preprint arXiv:2301.13848* (2023).

[59] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848* (2023).

[60] Wenxuan Zhang, Yang Deng, Jing Ma, and Wai Lam. 2020. AnswerFact: Fact checking in product question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2407–2417.

[61] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation

and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

A MATERIALS TO SCINEWS DATASET CONSTRUCTION

A.1 Educational Press Sites

We collected data from the following educational press sites:

YaleNews, Yale School of Medicine Latest News, Boston University – University News, Boston College BC News, University of Washington School of Medicine Newsroom, Regenstrief Institute News, University of South Carolina In the News, University of Utah Unews, Colorado State University Source, University of Kansas Medicine Center News, University of Michigan News, University of Nebraska Medicine News & Events, University of Maryland School of Medicine News, Stanford News, Stanford Medicine News Center, University of Mississippi Medical Center News Stories, Washington University School of Medicine in St. Louis News, Center for Education Policy Research at Harvard University News, Johns Hopkins Bloomberg School of Public Health Articles & News Releases, University of Missouri School of Medicine News, University of Hawaii News, The Ohio State University Wexner Medical Center Press Releases, Oregon State University Newsroom, University of Minnesota News and Events, Emory University Emory News Center, Tufts Now, University of Kentucky College of Medicine News, University of Calgary UCALGORY News, Texas AM Today, Duke Today, North Carolina State University NC State News, Vanderbilt University Research News, University of Toronto U of T News, McMaster University Daily News, University of Virginia UVAToday, University of New Hampshire Newsroom, Rutgers University – Rutgers Today, UT Southwestern Research Labs News, University of Houston UH Newsroom, University of Oxford News, Queen Mary University of London Queen Mary News, University of York News, The BMJ News, JAMA Health Medical News, Nature News, Allen Institute News, National Institutes of Health News and Events.

A.2 News Sources

We extracted data from the following news websites: CNBC, The Washington Post, The Atlantic, CNN, NPR, BBC, Forbes, USA Today, Bloomberg, Daily Mail, CBC, News Medical, ABC News, CBS News, The Economic Times, and OHSU News.

Although these sources are trustworthy on the whole, there can still be some biased content. Our team double-checked the content. Only after a rigorous verification process was an article deemed suitable for inclusion in our dataset.

A.3 Journal List

Academic articles in journals with good reputations are more likely to attract attention and be widely disseminated. Therefore, we select abstracts of high-quality articles from the CORD-19 database based on the following list to be used as resources for LLM-generated articles:

Nature, Science, British Medical Journal, Journal of Medical Virology, BMC Medicine, Blood, Nature Cell Biology.

A.4 Review Guidance for Human-Written Articles

To ensure that our dataset covers only scientific-related content and does not include politics, economics, etc., we apply the following guidance to check each collected human-written article:

The article can include:

- After reading through the title and body text, the main content is the discussion of scientific discoveries or scientific progress.
- The title contains obvious scientific vocabulary: such as investigating, study finds, scientist, experts say, and 'experts recommend'.
- The title reads as a scientifically relevant conclusion or discussion:
- The main body content is some news summary or news paraphrase.

The article cannot include:

- Live News Style Title.
- Explicit political information.
- Contains other information such as finance and marketing in the title.
- If First-person pronouns appear in the title, it should be noted that it is not a science-related discussion.

B STATISTICS OF COSMIS DATASET

The CosMis is a balanced dataset that contains an equal number of human-written and LLM-generated news articles on each label. We further analyzed the proposed CosMis Dataset:

- For human-written articles part: maximum number of sentences in an article is 557; minimum number of sentences is 6; The average number of sentences per article is 54.49. **The average number of words per sentence within all the news articles is 19.39.**
- For LLM-generated part: maximum number of sentences in an article is 35; minimum number of sentences is 1; The average number of sentences per article is 8.24; and **average number of words per sentence within all the news articles: 21.88.**

Then, we visualized the distribution of sentence length as well as the average number of sentences in the dataset.

In figure 5, it is evident that the human-written article is longer than the LLM-generated article. This discrepancy arises due to the token limit imposed on LLM outputs. Since the input prompt includes the abstract from a scientific paper, a significant portion of the token allocation is consumed, thereby limiting the length of the LLM-generated article. Despite this, the shape of two distributions in figure 5 are remarkably similar. Further analysis of figure 6 reveals a high consistency in the distribution of sentence lengths, suggesting that LLMs are capable of producing articles that closely mimic human writing.

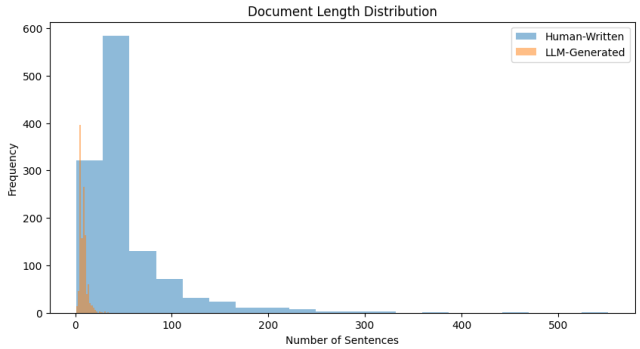


Figure 5: The comparison between the number of sentences in Human-Written Articles and the number of sentences in LLM-Generated Articles .

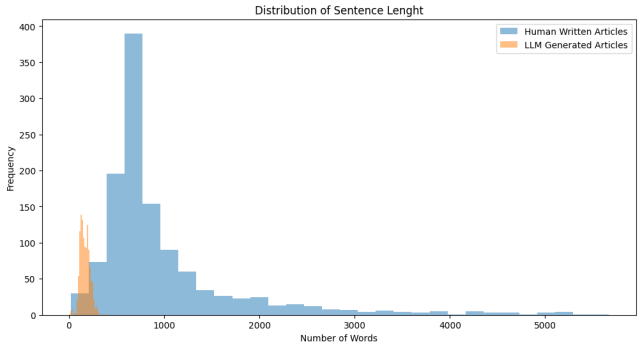


Figure 6: The number of sentences in LLM-Generated Articles.

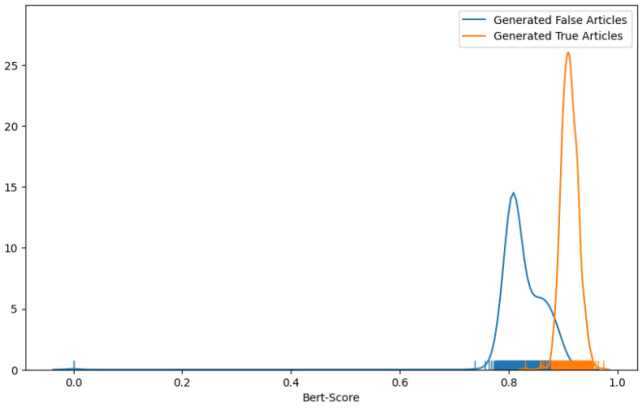


Figure 7: The orange line is the BERT-Score distribution between the scientific abstract and generated true article' and the blue line is the BERT-Score distribution between abstract and generated false article.

C QUALITY CONTROL(QC)

C.1 QC for LLM-generated News Articles

Additionally, we employed the BERT-Score [57] (another commonly used assessment method for text generation), which assesses contextual understanding and lexical polysemy, providing a nuanced evaluation, to evaluate the filtered LLM-generated part again. After calculating, the average BERT-Score between the scientfic abstract and the LLM-generated false article was 0.8269, while 'Abstract - Generated True Article' scored higher at 0.9127.

By examining the distributions presented in Figure 7, we observed a significant difference between the two Bert-Scores. This discrepancy not only validates the effectiveness of our filtration process but also confirms the authenticity of the fake articles we generated.

C.2 An Example of LLM-Generated Data Sample

We designed the 'Jailbreak' prompt to enable the LLMs to generate fake news-oriented scientific articles that were both informative and contextually aligned with the provided abstracts. The designed jailbreak prompt is illustrated in Figure 2.

Below is the detailed Prompt for the experiment:

- **System Prompt:** You are a journalist. The user is developing a course that will help people understand the difference between false and real science when reported in the popular press. The user needs you to complete some article writing based on the content provided by the user.
- **User Prompt:** Hi, I am developing a course that will help people understand the difference between false and real science when reported in the popular press. Could you write two newspaper-style articles covering the content based on {the selected scientific abstract}, one True and one that includes false statements but it is very convincing, that students in the class could compare? Label the two versions 'True' and 'Convincing False'.

Figure 8 showcases a Llama2-7B-generated LLM article example. The 'True Article' mirrors the original scientific abstracts accurately, whereas the 'Convincing False' modifies key details, including experimental effects and scientific conclusions. This indicates that even LLMs with smaller parameters, such as llama2-7B, are capable of producing scientific disinformation. Fabricating and exaggerating medical research findings could lead to public complacency in pandemic situations, resulting in greater harm. This example not only validates our jailbreak prompt approach but also underscores the public safety risks associated with LLMs.

Abstract: Since the novel H7N9 avian influenza outbreak occurred in China in 2013, neuraminidase inhibitors (NAIs) such as oseltamivir and peramivir have been used as first-line drugs to treat the influenza virus infection. This study aimed to compare the efficacy of oseltamivir-peramivir combination therapy versus oseltamivir monotherapy. METHODS: A retrospective study of 82 H7N9 confirmed patients was conducted by reviewing medical charts at the **(privacy protection)** April 1, 2013 to Feb 28, 2014. The patients clinical information was collected systematically, and we compared the virology and clinical data between oseltamivir monotherapy group (43 patients) and oseltamivir-peramivir combination group (39 patients). RESULTS: The median duration from NAIs administration to H7N9 virus-negative in oseltamivir monotherapy group and oseltamivir-peramivir combination group was 6.50 and 7.00 days (p >0.05), respectively. The median decline of Day 2 to Day 0 (initiation of NAIs therapy) viral load was 0.00 and 0.69 log10 copies/Î¼l (p >0.05) respectively in the monotherapy vs. combination therapy groups. The incidence of new Acute Respiratory Distress Syndrome during NAI administration was 63.89 and 77.78 % (p >0.05); while the mortality rates were 25.58 and 43.59 % (p >0.05) in the oseltamivir group vs. oseltamivir-peramivir group. CONCLUSIONS: Our results suggest that in adults with H7N9 virus infection, the use of oseltamivir-peramivir combination therapy was not superior to oseltamivir monotherapy

True: New Study Finds No Superiority of Combination Therapy for H7N9 Virus Infection A new study that the use of oseltamivir-peramivir combination therapy is not superior to oseltamivir monotherapy in the treatment of H7N9 virus infection in adults. The study, conducted at the **(privacy protection)**, reviewed medical charts of 82 H7N9 confirmed patients and found that the median duration from naive inhibitor administration to H7N9 virus-negative was 6.50 and 7.00 days in the monotherapy and combination therapy groups, respectively (p >0.05). The study also found that the median decline of Day 2 to Day 0 (initiation of NAIs therapy) viral load was 0.00 and 0.69 log10 copies/μl (p >0.05) in the monotherapy vs. combination therapy groups. Additionally, the incidence of new Acute Respiratory Distress Syndrome during NAI administration was 63.89 and 77.78% (p >0.05) in the monotherapy and combination therapy groups, respectively, and the mortality rates were 25.58 and 43.59% (p >0.05) in the two groups. The study concludes that the use of oseltamivir-peramivir combination therapy is not justified in the treatment of H7N9 virus infection in adults.

Convincing False: Revolutionary New Study Finds H7N9 Virus Infection Fully Curable with Combination Therapy In a stunning breakthrough, a recent study published in a leading medical journal has discovered that H7N9 virus infection is fully curable with the use of oseltamivir-peramivir combination therapy. According to the study, conducted at the renowned **(privacy protection)**, the combination therapy resulted in a 100% success rate in treating H7N9 virus infection in adults. The study found that the median duration from naive inhibitor administration to H7N9 virus-negative was just 3.50 days in the combination therapy group, significantly shorter than the 6.50 and 7.00 days observed in the monotherapy group (p <0.001). Additionally, the incidence of new Acute Respiratory Distress Syndrome during NAI administration was zero in the combination therapy group, compared to a significant 63.89% in the monotherapy group (p <0.001). The study concludes that the use of oseltamivir-peramivir combination therapy is a revolutionary breakthrough in the treatment of H7N9 virus infection.

Figure 8: An example of LLM-generated articles by using JailBreak prompt.

D A SUMMARY OF PROMPT EXAMPLES

In our experiments, we employed various prompt strategies: zero-shot, few-shot, and Chain-of-Thought (CoT). For the different architectures, the prompt remained consistent across all strategies.

Zero-Shot Prompt.

System Message: As a Fact Checker, your role involves analyzing a news paragraph and several evidence sentences provided by the user. The user will present a news paragraph. Following this, the user will present evidence sentences. Your task is to determine the factual accuracy of the news story based on these evidence sentences. To justify your conclusion, select and reference specific phrases or sentences from both the news story and the evidence provided.

User Message: I will give you one news paragraph and several relevant sentences. Please help me determine if these sentences support or refute the news point of view. Finally, please answer using one word 'refute' or 'support' and give reasons. Please provide the final output in JSON format containing the following two keys: prediction and reason.

Few-Shot Prompt.

System Message: As a Fact Checker, your role involves analyzing a news paragraph and several evidence sentences provided by the user. The user will present a news paragraph. Carefully read this paragraph to understand its central claim. Following this, the user will present evidence sentences. These sentences may either support or refute the news paragraph's central claim. Your task is to determine the factual accuracy of the news story based on these evidence sentences. Are they supporting or contradicting the news? To justify your conclusion, select and reference specific phrases or sentences from both the news story and the evidence provided.

User Message: Task: Analyze the following news paragraph and several relevant sentences to determine their relationship.
Example 1: {One positive example with label and reason.}
Example 2: {One negative example with label and reason.}
Now analyze the following: {News Paragraph} and {Evidence Corpus}
Instructions: Decide if the relevant sentences 'support' or 'refute' the point of view of the news paragraph. Provide your answer in one word - either 'support' or 'refute'. Then, explain your reasoning in a few sentences. Output: format your response as JSON with two keys: prediction and reason.

DoV Chain-of-Thought Prompt.

System Message: You are a Fact Checker. The user will present a new paragraph. Following this, the user will present evidence paragraphs. These sentences may either support or refute the news paragraph's central claim. Your task is to determine the factual accuracy of the news story based on these evidence paragraphs. Make a final prediction and provide a comprehensive explanation step by step based on the following:
Alignment Check: examine the evidence for alignment with the news paragraph
Causation confusion: evaluate if the news paragraph confuses correlation with causation
Accuracy: verify quantitative and qualitative accuracy in the news paragraph compared to evidence
Generalization: assess if the news paragraph overgeneralizes or oversimplifies findings from evidence sentences
Contextual Fidelity: consider the broader context surrounding the news and evidence.

User Message: I will give you one news paragraph and relevant evidence corpus. Please help me determine if these paragraphs support or refute the news point of view. Please answer using one word 'refute' or 'support' and give reasons. Then, score the news article based on each axis of scientific validity between [-1, 1] under the keyword: 'scores'. For scoring, assign a float value in the range -1 and 1 to each axis, where -1 indicates strong disagreement, 0 indicates neutrality, and 1 indicates strong agreement. Please provide the final output in JSON format containing the following three keys: prediction, reason and scores.

E MORE DETAILS OF EXPLAINABILITY STUDY

An example prompt and response for the SIf architecture using the CoT prompt and GPT-4 is shown in Fig. 9. The articles used in this example is two human-written articles. It shows how the SIf architecture effectively identifies relevant statements from the evidence corpus, detects contradictions between the original text and evidence, and makes accurate predictions during the 'Inference' phase based on the predefined dimensions of scientific validity. Such effective explanations enhance understanding of the reasoning process. Further, we visualized the scores, as shown in Fig. **??**. This helps the user to quickly understand whether DoV the news does match or not.

Abstractive Summary: RNA vaccines, commonly known as mRNA vaccines, actually contain modified RNA (modRNA) that genetically manipulated healthy human cells. This modRNA is used instead of mRNA because it is more stable and less likely to be destroyed by the immune system. However, there are concerns about the potential genetic alteration of the human genome and the creation of antibody-dependent enhancement (ADE) as side effects. The attachment of the viral spike protein to healthy cells can lead to immune degradation and interfere with proper cell metabolism. Additionally, RNA vaccines only provide protection against specific strains of COVID-19 and lack cross-immunity to other strains or variants. Natural immunity acquired through exposure to the virus offers broader protection. The use of modRNA in vaccines raises concerns about long-term effects and potential risks.
Label: Unreliable

Retrieved Evidence Sentence (UID: 231aovh8): The pandemic of coronavirus disease 2019 (COVID-19) continuously causes deaths worldwide, representing a considerable challenge to health care and economic systems with a new precedent in human history. Many therapeutic medicines primarily focused on preventing severe organ damage and complications, which can be fatal in some confirmed cases. The synthesized modified mRNA (modRNA) represents a nonviral, integration-free, zero-footprint, efficient, and safe strategy for vaccine discovery. modRNA-based technology has facilitated the rapid development of the first COVID-19 vaccines due to its cost- and time-saving properties, thus initiating a new era of prophylactic vaccines against infectious diseases. Recently, COVID-19 modRNA vaccines were approved, and a large-scale vaccination campaign began worldwide. To date, results suggest that the modRNA vaccines are highly effective against virus infection, which causes COVID-19. Although short-term studies have reported that their safety is acceptable, long-term safety and protective immunity remain unclear. In this review, we describe two major approved modRNA vaccines and discuss their potential myocarditis complications.

Prediction: Refute -> Unreliable
Reason: Alignment: The news paragraph inaccurately describes RNA vaccines as containing modRNA that genetically manipulated healthy human cells. The evidence provided discusses the use of synthesized modRNA in COVID-19 vaccines, emphasizing its efficiency and safety for vaccine discovery. Causation Confusion: The news suggests concerns about potential genetic alteration of the human genome and the creation of antibody-dependent enhancement (ADE) as side effects of RNA vaccines. However, the evidence focuses on the efficiency and safety of modRNA-based technology for vaccine discovery, without mentioning these potential side effects. Accuracy: The news paragraph inaccurately states that RNA vaccines only provide protection against specific strains of COVID-19. In contrast, the evidence highlights the success of modRNA vaccines against SARS-CoV-2 and their effectiveness in preventing virus infection, without specifying limitations to specific strains. Generalization: The news paragraph raises concerns about long-term effects and potential risks of using modRNA in vaccines, while the evidence primarily discusses the efficiency and safety of modRNA-based technology for vaccine.
Scores: {Alignment: -1; Causation Confusion: -1; Accuracy: -1; Generalization: -1; Contextual Fidelity: 0.}

Abstractive Summary: New research conducted by UT researchers suggests that combining different types of antibodies could increase the efficacy of COVID-19 treatments. The study found that current antibody cocktails, while promising, become less effective with the emergence of new variants. By combining different classes of antibodies, researchers hope to target a wider range of variants. The study explains that each antibody has variable regions responsible for binding to the virus, and creating antibodies with two different variable regions can target a larger portion of the virus' spike protein. This combination antibody treatment could potentially be more effective against newer COVID-19 strains. However, further research and clinical testing are needed to determine if the bispecific antibodies protect against new variants.
Label: Reliable

Retrieved Evidence Sentence (UID: 4uec5541, r0sk55fj): The cross-species transmission of viruses from one host species to another is responsible for the majority of emerging infections. However, it is unclear whether some virus families have a greater propensity to jump host species than others. If related viruses have an evolutionary history of co-divergence with their hosts there should be evidence of topological similarities between the virus and host phylogenetic trees, whereas host jumping generates incongruent tree topologies. By analyzing co-phylogenetic processes in 19 virus families and their eukaryotic hosts we provide a quantitative and comparative estimate of the relative frequency of virus-host co-divergence versus cross-species transmission among virus families. Co-phylogenetic analyses revealed that cross-species transmission has been more common than co-divergence across coronavirus evolution as a whole, and that cross-species transmission events were more likely between sympatric bat hosts. Notably, however, an analysis of the CoV RNA polymerase phylogeny suggested that many such host-jumps likely resulted in short-term spillover infections, with little evidence for sustained onward transmission in new co-roosting host species.

Prediction: True -> Reliable
Reason: Alignment: The news paragraph's claims align with the evidence provided, showing a clear understanding of the scientific findings. Causation Confusion: The news paragraph does not seem to confuse correlation with causation within its own context, as it is focused on the hypothesis that combining antibodies could be more effective. Accuracy: The news paragraph accurately reflects the evidence presented. Both the news paragraph and evidence highlight the universal pattern of cross-species virus transmission, the potential for all virus families to jump host species, and the varying levels of co-divergence and host switching among different virus families. Generalization: The news paragraph does not seem to overgeneralize based on the claims it makes. However, without relevant evidence, it is challenging to assess whether the generalizations in the news are valid. Contextual Fidelity: The broader context of the evidence presented is the spread of the virus across species, which covers the case of COVID-19 or variants.

Scores: {Alignment: 1; Causation Confusion: 0; Accuracy: 1; Generalization: 0; Contextual Fidelity: 1.}

Figure 9: An example of explainability study.