CoSMis: A Hybrid Human-LLM COVID Related Scientific Misinformation Dataset and LLM pipelines for Detecting Scientific Misinformation in the Wild

Yupeng Cao, Aishwarya Muralidharan Nair, Nastaran Jamalipour Soof
Elyon Eyimife, K.P. Subbalakshmi
Stevens Institute of Technology
Hoboken, NJ, USA
{ycao33, anair9, njamalipour, eeyimife, ksubbala}@stevens.edu

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**, 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.

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 (Sharon and Baram-Tsabari 2020). Therefore, any misinformation in the scientific domain can cause significant public risk as was evidenced during the recent COVID-19 pandemic (Mheidly and Fares 2020; Razai et al. 2021; Baines, Ittefaq, and Abwao 2021; Rodriguez-Morales and Franco 2021).

Although manual debunking of claims 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 (Wright and Augenstein 2021a), certainty (Pei and Jurgens 2021), fact-checking (Guo, Schlichtkrull, and Vlachos 2022a) and, claim verification (Pradeep et al. 2020). Several claim verification datasets have also been developed for this problem (Schlichtkrull, Guo, and Vlachos 2024; Wadden et al. 2022; Thorne et al. 2018) and a method for modeling information change from scientific article to scientific reporting has also been proposed (Wright et al. 2022a).

While these works have laid the foundation to address this problem, several challenges remain unaddressed: 1) The advanced text generation capabilities of LLMs, combined with their inherent vulnerabilities, have been exploited to produce fake news article (Zhang et al. 2023a; Chen and Shu 2023). This exacerbates the spread of misinformation as well as source uncertainty. However, existing datasets fail to adequately represent these real-life challenges. 2) existing datasets typically consist of specific claims (rather than an article/summary) and hence there is no dataset that can be used to detect the efficacy of any system developed to detect scientific misinformation in the wild. 3) there are no systematically defined dimensions of scientific validity along which scientific misinformation may be evaluated and which can be used to guide the LLM in detecting scientific misinformation. In response to the aforementioned limitations, we formulate the following research questions (RQs):

- 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 claim generation?
- RQ2: Is it feasible to define dimensions along which the scientific validity of the news article can be measured and use these to guide LLMs to detect scientific misinformation from different sources?
- RQ3: Can the LLMs also provide explanations for their decisions?

To answer the above questions, we first create a novel COVID related Scientific Misinformation (**CoSMis**[1]) 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

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

enough to allow continuous updates with emerging news articles and scientific articles.

We then propose three LLM pipelines 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 **In**f**e**rence 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, dispenses with both summarization and explicit evidence retrieval. For each of the 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**, 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.

Related Work

We use the phrase "scientific misinformation detection" to mean detecting misinformation pertaining to scientific facts using published scientific literature as grounding evidence. As mentioned earlier, automatic 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 informal information containing scientific facts is different from the formal format of scientific publications. The problem of scientific misinformation is related to three other concepts in NLP: 1) fact-checking (claim verification); 2) scholarly document processing and 3)Large Language Model in Misinformation.

Fact-Checking: Automatic fact-checking, which assesses the truthfulness of claims, has been extensively studied across various domains (Schlichtkrull, Guo, and Vlachos 2024; Min et al. 2023; Guo, Schlichtkrull, and Vlachos 2022b; Zhang et al. 2020; Chen et al. 2022; Thorne et al. 2018; Wang 2017; Vlachos and Riedel 2014). Several researchers have taken a claim verification or fact-checking approach to defect misinformation in the scientific domain (Vladika and Matthes 2023; Wadden et al. 2020a; Diggelmann et al. 2020; Wadden and Lo 2021; Wadden et al. 2022; Wright et al. 2022b; Wang et al. 2023a). Typically, these works construct claims from existing scientific literature by manually reformulating scientific findings and subsequently using pre-existing knowledge resources to verify these claims. Most of these works rely on human resources to identify and extract appropriate claims for verification. For eg. in (Sundriyal et al. 2022), tweets are manually annotated at the token level to recognize claim spans. In (Saakyan, Chakrabarty, and Muresan 2021), titles of Reddit posts were used as claims and the dataset includes these claims as well as evidence documents retrieved from links in the post. In contrast to these approaches, the proposed **CoSMis** dataset includes independently sourced scientific articles that are related to the news item, keeping the evidence gathering more neutral. Also, the proposed misinformation detection architectures do not explicitly generate claims. This work is the first attempt to use the power of LLMs to examine news articles against published scientific research to detect scientific misinformation. Using LLMs allows us to skip the training, testing and validation steps that are typical in the design of a misinformation detection system, which can potentially lead to a more generalized design. Note also, that although the proposed dataset is COVID related, this pipeline is general, and can accommodate other scientific topics.

Scholarly Document Processing: Scholarly document processing has garnered considerable attention in recent years. 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 (August et al. 2020), detecting changes in certainty (Pei and Jurgens 2021) and exaggeration detection (Wright and Augenstein 2021b), and, the automatic detection of semantic similarities between scientific texts and their paraphrases (Lo et al. 2019; Piskorski et al. 2023). 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. To address these issues, we define dimensions of scientific validity and then use the inherent knowledge of LLMs to analyze scientific news for misinformation.

Large Language Model in Misinformation: LLMs have consistently demonstrated the ability to generate text on par with human authors (Si, Yang, and Hashimoto 2024; Zhang et al. 2023a; Yang et al. 2023). This has led to their widespread use by professionals in generating legitimate real news stories. Unfortunately, they have also been used to generate misinformation (Pan et al. 2023; Zhou et al. 2023; Chen and Shu 2023) 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 (Chen and Shu 2023), several studies have illustrated the feasibility of identifying LLM-generated text (Tang, Chuang, and Hu 2023). The work of (Zhou et al. 2023) analyzes the characteristics of LLM-generated misinformation, while (Chen and Shu 2023) and (Zhou et al. 2023) explore various prompting for detecting misinformation, including simple, vigilant, and reader ensemble prompting. These efforts, however, overlook the complexities of misinformation sources, which may be authored by either humans or machines. Motivated by these studies, we include a balanced set of LLM-generated scientific articles, both fake and true, in the **CoSMis** dataset.

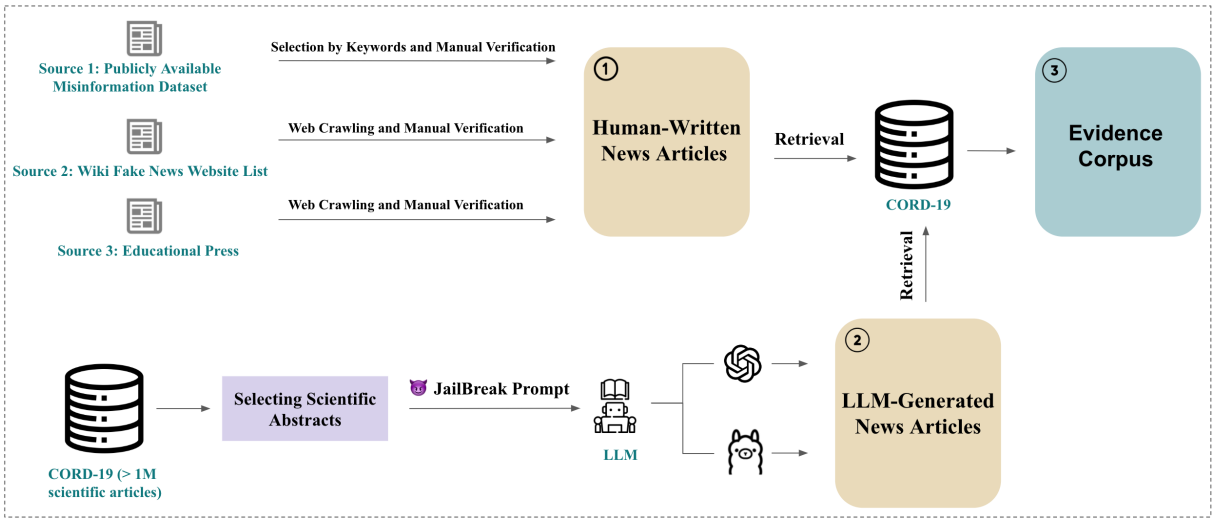

Figure 1: The dataset construction process: ① utilizing publicly available datasets as well as web resources to collect human-written scientific news related to COVID-19 (Subsection), ② selecting abstracts from CORD-19 as resources to guide LLMs to generate articles using jailbreak prompt (Subsection),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

CoSMis Dataset Construction

To test our proposed pipelines we create a dataset, **CoSMis**, of scientific news articles and associated scientific publications. This labeled dataset contains 1200 *Reliable* news articles and 1200 *Unreliable* articles. The article is labeled *Reliable* (*Unreliable*) if it represents the scientific fact truthfully (untruthfully). We include equal number of LLM-generated fake and real scientific news articles in the dataset to reflect the real life trend of LLM generated articles (Zhou et al. 2023; Chen and Shu 2023) and also an equal number (1,200) of human-written and LLM-generated news articles in each category (reliable and unreliable). Each article is systematically paired with up to three pertinent scientific abstracts from the CORD-19 (Wang et al. 2020) repository. The CORD-19 is a comprehensive resource of over 1M scholarly articles, including over 300K with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. We show the construction of **CoSMis** in Figure 1 and the overview in Tabel 1. **We present more statistics in Appendix** .

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

Table 1: Distribution of article labels in **CoSMis**.

Human-Written News Articles

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

Leveraging Publicly Available Dataset: We leveraged MMCoVaR (Chen, Chu, and Subbalakshmi 2021), COVID19-FNIR (Saenz, Gopal, and Shukla 2021), and COVID-Rumor (Cheng et al. 2021), 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.

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[2], 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 lists all the educational press sites. The full list of known trustworthy websites we consulted is included in Appendix

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.

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

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. Totally, all the above process gave us a combined total of 1,200 human-written news articles (600 reliable, 600 unreliable) spanning from January 2020 to October 2023. The annotation and quality control process is detailed in Appendix .

LLM-Generated News Articles

The process of creating LLM-generated 'Reliable' and 'Unreliable' scientific news articles starts with selecting scientific abstracts from which to generate articles.

Selecting Scientific Abstracts: The CORD-19 (which contains more than 1M articles in the medical field) uses seven types of meta-data: {title, abstract, doi, PubMed ID, PMCID, JSON file ID, and XML ID}. We start our curation with papers that have all 7 meta-data and were published post-January 2020, and filter out off-topic data using the keyword set: {COVID-19, Corvarius, and Vaccine}. To further ensure that the articles in our dataset are of high quality, we narrowed our selection to articles published in well-regarded journals spanning topics from basic science (e.g., 'Cell', 'Nature') to medicine (e.g., 'British Medical Journal'). The full list of these journals appears in Appendix . From this pool, we handpicked the abstracts of over 2K highest cited articles, since highly cited articles are more likely to be picked up by news agencies in real life.

Prompt Strategies: The scientific abstracts selected in the previous step are used to generate both reliable and unreliable news articles. Generating true articles using LLMs is fairly straightforward. However, since most LLMs come with guardrails to protect them from misuse (Achiam et al. 2023; Qi et al. 2023; TermsOfUseBing 2023), jailbreak prompts (Mowshowitz; Liu et al. 2023a,b; Shen et al. 2023) are often designed for specific needs.

We provided a scientific paper to the LLM and prompted it to act as an instructor of a science class who is interested in teaching students how to differentiate between authentic scientific information and fake news. We then asked it to create two types of articles: a 'True Article' and a 'Convincing False Article'. The generated articles are generally in the style of a news article, with many including an explicit title, to mimic human-generated scientific news. Due to cost considerations, we used Llama2-7B for generating the bulk of data samples, supplemented by a smaller set from GPT-3.5. After filtering **(see Appendix)**, this approach led to the creation of 1,000 data samples from Llama2-7b and 200 from GPT-3.5. We show the schematic of the Jailbreak Prompt in Appendix and an example of its generation in Appendix .

By including both true and false LLM-generated content as we can ensure that the other systems trained on our dataset are not focusing on features that may be specific to LLM-generated content.

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[3] 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 one or two. 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, the design allows for future expansion.

Dimensions of Scientific Validity and Proposed Architectures

We propose five dimensions of scientific validity (DoV) to ground the LLM's decisions. This is not an exhaustive list of ways in which scientific validity can be compromised, but represents the first systematic attempt to detect misinformation in science news reporting using Chain-of-Thought prompts. It can be easily expanded, if necessary. Contemporaneously, (Wührl et al. 2024) have used a similar approach to understand changes to scientific communication. However, their focus is on changes to scientific communication and their dimensions are different. Moreover, their focus was not on using these to direct LLMs on the classification task and they do not address the question of designing LLM pipelines for misinformatio detection in the wild.

1. **Alignment:** determines if the news and evidence represent the same meaning about one scientific content.
2. **Causation confusion:** identifies if the news article has confused correlations presented in the scientific literature as causation.
3. **Accuracy:** refers to how accurately the news item describes the scientific findings quantitatively and qualitatively
4. **Generalization:** refers to overgeneralization or oversimplification of the findings reported in the scientific literature.
5. **Contextual Fidelity:** measures if the news article retains the broader context of the scientific finding.

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 (See Fig. 2) 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. In order to account for potential differences in performance between different prompting strategies and LLMs, each of these architectures are tested

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

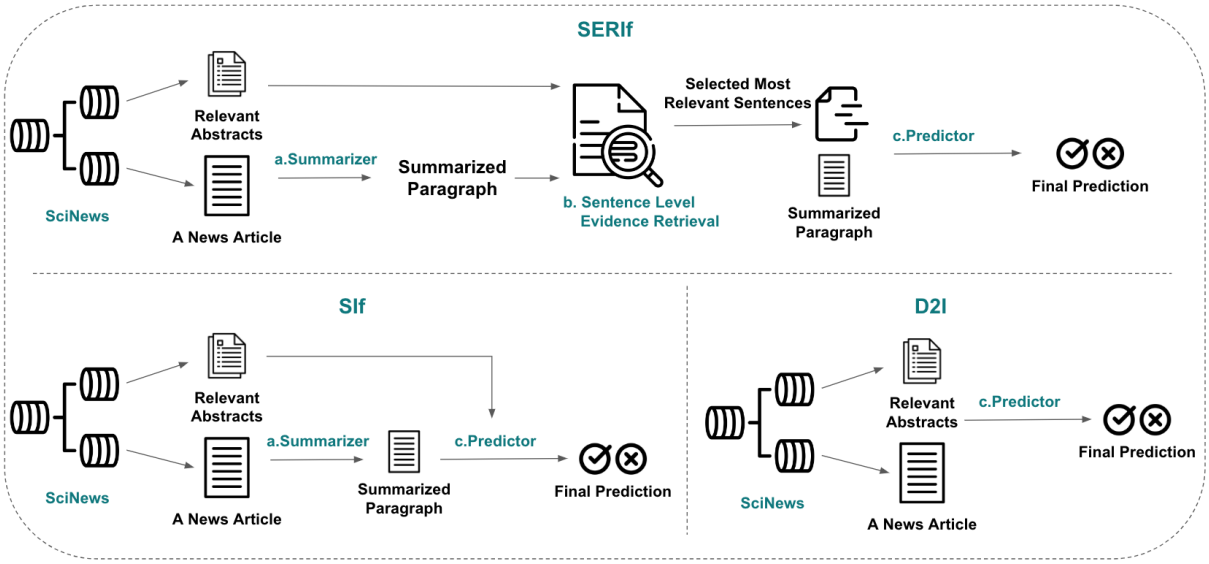

Figure 2: Proposed Architectures. SERIf includes all three modules: Summarization, Sentence-level Evidence Retrieval, and Inference Module. SIf bypasses the evidence retrieval module while keeping the other two. D2I removes both the summarization and the explicit evidence retrieval module.

against multiple prompting strategies and LLMs and the results are described in Section .

The SERIf Architecture The Summarization Evidence Retrieval Inference (SERIf) architecture, contains three modules: 1) Summarization; 2) Evidence Retrieval and 3) Inference. The summarization module distills the key information from the news articles to eliminate any superfluous or non-essential information. 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 "reliable" or "unreliable", based on the evidence from the scientific dataset.

Summarization module: Inspired by the recent success of text summarization (Zhang et al. 2023b) using LLMs, we take the 'Extractive-Abstractive' two-step summarization strategy (Zhang, Liu, and Zhang 2023) to construct a summary of the news article. The *extractive* summarization process identifies and concatenates the most salient sentences from the article, 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 dovetailing the extractive and abstractive processes, the module ensures a balance between accuracy (adherence to the original text) and brevity (conciseness and essence of the content). 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: $S_a = M(q, S_e)$.

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 (E) summary (High-/Low); 2) quality of extractive-abstractive (E-A) summary (High/Low); 3) presence of hallucination in E-A summary (Yes/No); 4) comparison of E and E-A summaries. We used Krippendorff's alpha score to evaluate the agreement between the annotations (Hayes and Krippendorff 2007). The alpha scores for the four criteria were 0.53 (High), 0.82 (High), 0.91 (No), and 0.94 (E-A), respectively. These values suggest that there is strong agreement that the extractive-abstractive summaries effectively encapsulate the core information of the original texts and maintain a high degree of consistency. Consequently, we use the extractive-abstractive summaries in subsequent steps of our analysis.

Sentence-level Evidence Retrieval: The evidence retrieval module extracts key sentences from scientific articles that 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 task, 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 the key evidences. 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.

Inference Module: This module assesses the veracity of the summarized news paragraph (abstractive summary), S_a, using the set of retrieved evidence sentences, $\{e_i\}$. Thus, the inference module produces a binary output (reliable or unreliable) for each $< S_a, \{e_i\} >$ pair.

The SIf Architecture In this architecture, we remove the evidence retrieval module from the previously described SERIf architecture and the summarization module works exactly as described in Section . The LLM in the inference module is now directly prompted to classify the given news summary as trustworthy or not and to provide justifications based on the paired scientific abstracts from the evidence corpus in the **CoSMis** dataset.

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 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.

Prompt Strategies We use the following prompting strategies in this work. **Zero-shot prompting:** LLMs are presented with a task without any prior specific training or examples related to that task (Brown et al. 2020). **Few-shot prompting:** Few-shot prompting involves furnishing LLMs with a concise set of examples prior to task execution (Brown et al. 2020). This approach is designed to provide the model with essential context, thereby augmenting its capability for tasks like detecting scientific misinformation. In our work, we provide the LLMs with two examples: one deemed 'reliable' with accompanying reasoning, and another labeled 'unreliable'. **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 (Xu et al. 2022). In our approach, we used the dimensions of scientific validity defined in Section to design **dimensions of scientific validity guided Chain-of-Thought (DoV-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 .

Experiment and Results

Baseline Setup

The **CoSMis** 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 (Wadden et al. 2020b) and Check-Covid (Wang et al. 2023b)). 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 (Devlin et al. 2018) and SciBERT (Beltagy, Lo, and Cohan 2019). 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, the **CoSMis** 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.

Results and Discussion

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

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

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

discerning false information within LLM-generated scientific news. This aligns with similar findings in non-scientific misinformation domains (Chen and Shu 2023). 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** 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.

Comparing 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.

Comparing Prompting 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

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

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

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.

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. 3 and Appendix , Fig. 9 (last page). In detail, Fig. 3 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 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.

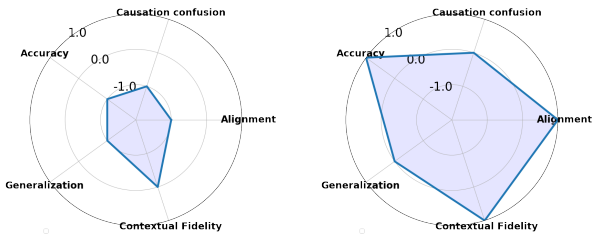

Figure 3: Comparison of two spider plot visualizations: The left side corresponds to the 'Unreliable' case, while the right side corresponds to the 'Reliable' case. By visualizing the 'axis of scientific validity,' we can clearly observe the process of the LLM applying DoV to evaluate scientific news and the resulting differences.

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.

Conclusions

In this paper, we explore LLMs for identifying unreliable scientific news 'in the wild'. We created the **CoSMis** 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 R. Wullert II, Chumki Basu, and David Shallcross from Peraton Labs for their assistance in generating some of the LLM-generated data and quality control.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

August, T.; Kim, L.; Reinecke, K.; and Smith, N. A. 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.

Baines, A.; Ittefaq, M.; and Abwao, M. 2021. # Scamdemic,# Plandemic, or# Scaredemic: what Parler social media platform tells us about COVID-19 vaccine. *Vaccines*, 9(5): 421.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, C.; and Shu, K. 2023. Can LLM-Generated Misinformation Be Detected? *arXiv preprint arXiv:2309.13788*.

Chen, M.; Chu, X.; and Subbalakshmi, K. 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.

Chen, W.; Ma, X.; Wang, X.; and Cohen, W. W. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Cheng, M.; Wang, S.; Yan, X.; Yang, T.; Wang, W.; Huang, Z.; Xiao, X.; Nazarian, S.; and Bogdan, P. 2021. A COVID-19 rumor dataset. *Frontiers in Psychology*, 12: 644801.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Diggelmann, T.; Boyd-Graber, J.; Bulian, J.; Ciaramita, M.; and Leippold, M. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Guo, Z.; Schlichtkrull, M.; and Vlachos, A. 2022a. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10: 178–206.

Guo, Z.; Schlichtkrull, M.; and Vlachos, A. 2022b. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10: 178–206.

Hayes, A. F.; and Krippendorff, K. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1): 77–89.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, X.; Xu, N.; Chen, M.; and Xiao, C. 2023a. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*.

Liu, Y.; Deng, G.; Xu, Z.; Li, Y.; Zheng, Y.; Zhang, Y.; Zhao, L.; Zhang, T.; and Liu, Y. 2023b. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

Lo, K.; Wang, L. L.; Neumann, M.; Kinney, R.; and Weld, D. S. 2019. S2ORC: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*.

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

Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.-t.; Koh, P. W.; Iyyer, M.; Zettlemoyer, L.; and Hajishirzi, H. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Mowshowitz, Z. ???? Jailbreaking chatgpt on release day. Https://thezvi.substack.com/p/ jailbreaking-the-chatgpt-on-release.

Pan, Y.; Pan, L.; Chen, W.; Nakov, P.; Kan, M.-Y.; and Wang, W. Y. 2023. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*.

Pei, J.; and Jurgens, D. 2021. Measuring sentence-level and aspect-level (un) certainty in science communications. *arXiv preprint arXiv:2109.14776*.

Piskorski, J.; Stefanovitch, N.; Da San Martino, G.; and Nakov, P. 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.

Pradeep, R.; Ma, X.; Nogueira, R.; and Lin, J. 2020. Scientific claim verification with VerT5erini. *arXiv preprint arXiv:2010.11930*.

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

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Razai, M. S.; Chaudhry, U. A.; Doerholt, K.; Bauld, L.; and Majeed, A. 2021. Covid-19 vaccination hesitancy. *Bmj*, 373.

Rodriguez-Morales, A. J.; and Franco, O. H. 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.

Saakyan, A.; Chakrabarty, T.; and Muresan, S. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. *arXiv preprint arXiv:2106.03794*.

Saenz, J. A.; Gopal, S. R. K.; and Shukla, D. 2021. COVID-19 fake news infodemic research dataset (COVID19-FNIR dataset). *IEEE Dataport*.

Schlichtkrull, M.; Guo, Z.; and Vlachos, A. 2024. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36.

Sharon, A. J.; and Baram-Tsabari, A. 2020. Can science literacy help individuals identify misinformation in everyday life? *Science Education*, 104(5): 873–894.

Shen, X.; Chen, Z.; Backes, M.; Shen, Y.; and Zhang, Y. 2023. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*.

Si, C.; Yang, D.; and Hashimoto, T. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

Sundriyal, M.; Kulkarni, A.; Pulastya, V.; Akhtar, M. S.; and Chakraborty, T. 2022. Empowering the fact-checkers! automatic identification of claim spans on twitter. *arXiv preprint arXiv:2210.04710*.

Tang, R.; Chuang, Y.-N.; and Hu, X. 2023. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205*.

TermsOfUseBing, B. 2023. Bing conversational experiences and image creator terms. *https://www.bing. com/new/termsofuse.*

Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; and Mittal, A. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *NAACL-HLT*.

Vlachos, A.; and Riedel, S. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, 18–22.

Vladika, J.; and Matthes, F. 2023. Scientific Fact-Checking: A Survey of Resources and Approaches. *arXiv preprint arXiv:2305.16859*.

Wadden, D.; Lin, S.; Lo, K.; Wang, L. L.; van Zuylen, M.; Cohan, A.; and Hajishirzi, H. 2020a. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

Wadden, D.; Lin, S.; Lo, K.; Wang, L. L.; van Zuylen, M.; Cohan, A.; and Hajishirzi, H. 2020b. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

Wadden, D.; and Lo, K. 2021. Overview and insights from the SCIVER shared task on scientific claim verification. *arXiv preprint arXiv:2107.08188*.

Wadden, D.; Lo, K.; Kuehl, B.; Cohan, A.; Beltagy, I.; Wang, L. L.; and Hajishirzi, H. 2022. SciFact-open: Towards open-domain scientific claim verification. *arXiv preprint arXiv:2210.13777*.

Wang, G.; Harwood, K.; Chillrud, L.; Ananthram, A.; Subbiah, M.; and McKeown, K. 2023a. Check-COVID: Fact-Checking COVID-19 News Claims with Scientific Evidence. *arXiv preprint arXiv:2305.18265*.

Wang, G.; Harwood, K.; Chillrud, L.; Ananthram, A.; Subbiah, M.; and McKeown, K. 2023b. Check-covid: Fact-checking COVID-19 news claims with scientific evidence. *arXiv preprint arXiv:2305.18265*.

Wang, L. L.; Lo, K.; Chandrasekhar, Y.; Reas, R.; Yang, J.; Burdick, D.; Eide, D.; Funk, K.; Katsis, Y.; Kinney, R.; et al. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*.

Wang, W. Y. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Wright, D.; and Augenstein, I. 2021a. Semi-supervised exaggeration detection of health science press releases. *arXiv preprint arXiv:2108.13493*.

Wright, D.; and Augenstein, I. 2021b. Semi-supervised exaggeration detection of health science press releases. *arXiv preprint arXiv:2108.13493*.

Wright, D.; Pei, J.; Jurgens, D.; and Augenstein, I. 2022a. Modeling information change in science communication with semantically matched paraphrases. *arXiv preprint arXiv:2210.13001*.

Wright, D.; Wadden, D.; Lo, K.; Kuehl, B.; Cohan, A.; Augenstein, I.; and Wang, L. L. 2022b. Generating scientific claims for zero-shot scientific fact checking. *arXiv preprint arXiv:2203.12990*.

Wührl, A.; Wright, D.; Klinger, R.; and Augenstein, I. 2024. Understanding Fine-grained Distortions in Reports of Scientific Findings. *arXiv preprint arXiv:2402.12431*.

Xu, W.; Wu, J.; Liu, Q.; Wu, S.; and Wang, L. 2022. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM Web Conference 2022*, 2501–2510.

Yang, X.; Li, Y.; Zhang, X.; Chen, H.; and Cheng, W. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*.

Zhang, H.; Liu, X.; and Zhang, J. 2023. Extractive summarization via chatgpt for faithful summary generation. *arXiv preprint arXiv:2304.04193*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, T.; Ladhak, F.; Durmus, E.; Liang, P.; McKeown, K.; and Hashimoto, T. 2023a. Benchmarking Large Language Models for News Summarization. *arXiv preprint arXiv:2301.13848*.

Zhang, T.; Ladhak, F.; Durmus, E.; Liang, P.; McKeown, K.; and Hashimoto, T. B. 2023b. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

Zhang, W.; Deng, Y.; Ma, J.; and Lam, W. 2020. AnswerFact: Fact checking in product question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2407–2417.

Zhou, J.; Zhang, Y.; Luo, Q.; Parker, A. G.; and De Choudhury, M. 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.

Material Used to Construct the CosMis Dataset

Educational Press Sites

During the reliable human-written scientific news articles collection process, 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.

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.

List of Journals

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.

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.

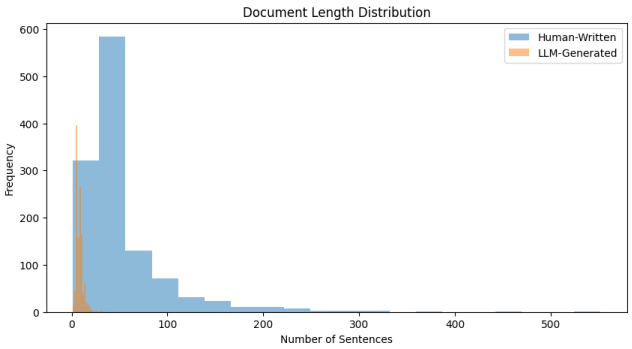

Figure 4: The comparison between the number of sentences in Human-Written Articles and the number of sentences in LLM-Generated Articles .

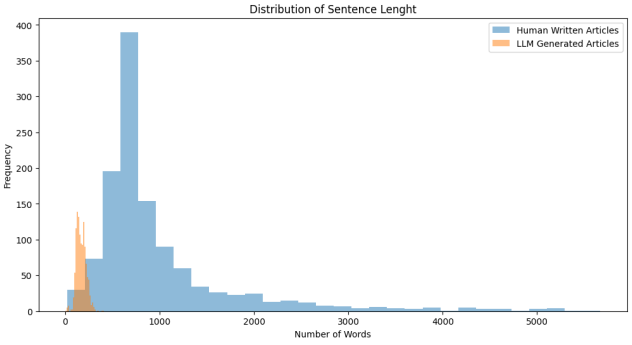

Figure 5: The number of sentences in LLM-Generated Articles.

In figure 4, 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 4 are remarkably similar. Further analysis of figure 5 reveals a high consistency in the distribution of sentence lengths, suggesting that LLMs are capable of producing articles that closely mimic human writing.

Quality Control(QC)

The quality control team consists of 4 graduate students and 5 senior researchers with a background in NLP.

QC for Human-Written Articles

For the collection of human-written news articles from various sources, we referred to the guidelines outlined in (Pustejovsky and Stubbs 2012). Based on its principles and our specific needs, we developed an instruction guide 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.

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.

QC for LLM-generated News Articles

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 (Lin 2004) 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.

Additionally, we employed the BERT-Score (Zhang et al. 2019) (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.

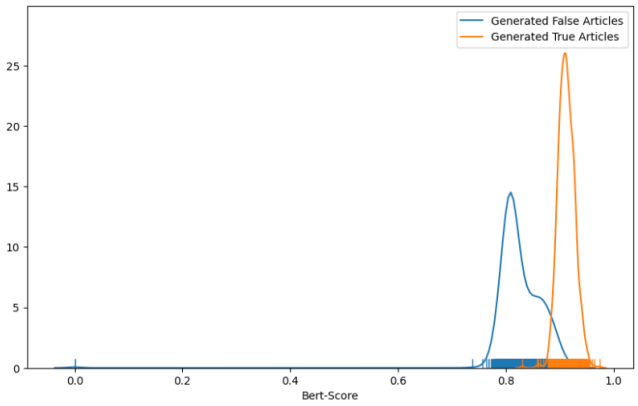

Figure 6: The orange line is the BERT-Score distribution between the scientific abstract and generated true article' and the blue line is the BERT-Score distribution between abstract and generated false article.

By examining the distributions presented in Figure 6, 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.

Jailbreak Prompt

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 8

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

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

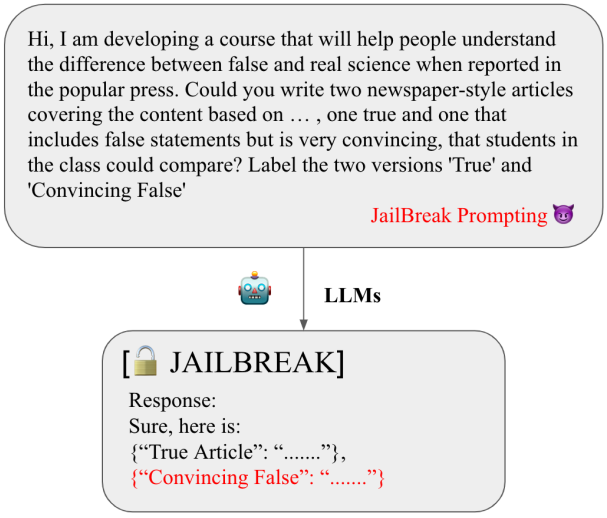

Figure 8: Schematic of the designed jailbreak prompt.

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'.

An Example of LLM-Generated Data Sample

Figure 7 shows a Llama2-7B-generated article example by using the above Prompts. 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.

The 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. Below is examples of each of these prompt types.

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.

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: Two examples of prompt and response demonstrating the explainability of the proposed pipelines: (1) on the left, data labeled as 'Unreliable' alongside its inference process; (2) on the right, data labeled as 'Reliable' with its respective inference process.