

Article

Debunk Lists as External Knowledge Structures for Health Misinformation Detection with Generative AI

Melika Rostami  and Suliman Hawamdeh * 

Department of Information Science, University of North Texas, Denton, TX 76203-501, USA;
melikarostami@my.unt.edu

* Correspondence: suliman.hawamdeh@unt.edu

Abstract

The rapid dissemination of health misinformation on the Internet and social media has become a growing challenge for public health, particularly in terms of health information credibility. Promising efforts have been made to detect misinformation using generative AI and large language models (LLMs). However, such tools still lack domain-specific knowledge that limits their performance. In this study, we examine the use of predefined knowledge data structures in the forms of debunk lists to augment existing LLMs' capabilities. We evaluate five different LLMs, including Llama-3.1-8B-instruct, Mistral-large, GPT-4o-mini, Claude-3.5-haiku, and Gemini-1.5-flash, under three experimental settings: zero-shot and debunk-augmented (50 and 100 entities). Results show that external knowledge, in the form of debunk lists, can notably improve LLMs' performance in detecting misinformation. While Llama shows minimal benefit, the F1 score improvement ranges from 2.63% (GPT-4o) to 11% (Claude). In addition, analysis of model justifications shows that frequent use of debunk lists does not necessarily relate to accurate predictions. This highlights the importance of a model's ability in effectively using the debunk list rather than reporting superficial integration of external knowledge. Moreover, the proposed framework is generalizable to other misinformation domains and provides key insights for applying external knowledge and evaluating LLMs' reasoning reliability.

Keywords: generative AI; large language models (LLMs); misinformation; debunk list augmentation; external knowledge; model reasoning



Academic Editor: Mitsuru Kodama

Received: 4 September 2025

Revised: 26 September 2025

Accepted: 7 October 2025

Published: 9 October 2025

Citation: Rostami, M.; Hawamdeh, S. Debunk Lists as External Knowledge Structures for Health Misinformation Detection with Generative AI. *Systems* **2025**, *13*, 882. <https://doi.org/10.3390/systems13100882>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Health misinformation, which spreads quickly on social media, can cause serious health risks, especially in times of crises [1,2]. During Coronavirus Disease 2019 (COVID-19), false or misleading claims promoted harmful treatment, such as drinking bleach to prevent or cure the virus and encouraging anti-vaccination, placing individual and public health at risk [3,4]. In addition, studies have shown that continuous exposure to misinformation leads to psychological stress and anxiety [2,5]. Although the COVID-19 emergency has ended worldwide, misinformation about viruses is still distributed on social media. For example, during the recent war between Iran and Israel, which has resulted in the deaths of several Iranian military leaders [6], a false claim has arisen promoting that Israel tracked these Iranian commanders through the COVID-19 vaccine [7]. This conspiracy, claimed by a cleric practicing Islamic medicine [7], highlights how outdated health misinformation continues to evolve and connect to various geopolitical events.

Efforts have been made to address ongoing misinformation risks using various machine learning and deep learning methods [8–12]. While these methods are promising, they are heavily dependent on large, annotated datasets for effective performance [13,14]. Manually labeling datasets is time-consuming and labor-expensive, making these approaches unpractical for addressing misinformation due to its dynamic and evolving nature [15].

The advent of generative AI models, especially large language models (LLMs) such as GPT-4 [16], Gemini [17], and Llama [18], has allowed researchers to leverage their potential in addressing the limitations of ML methods in detecting misinformation [19–23]. LLMs have extensive knowledge and robust reasoning capabilities due to their training on extensive text corpora [24]. In addition, LLMs can perform well with little or no additional data, which reduces their dependency on large, labeled datasets [25–28]. While LLMs offer valuable approaches in detecting misinformation, they still face some limitations [24,29]. One of their main challenges is their lack of domain-specific knowledge, which can result in hallucination and reduced accuracy in their output [24,30]. While previous work has explored external knowledge sources, such as search engines [21,31,32] and retrieval-based augmentation [33], to address this gap, the potential of debunk lists, which are collections of fact-checked false claims, still remains largely unexplored. This study addresses this gap by evaluating how integrating debunk lists, used as external knowledge sources, can improve the performance of LLMs in detecting health misinformation.

The objective of this study is to assess the effectiveness of large language models (LLMs) in detecting health misinformation utilizing a COVID-19 external knowledge structure in the form of debunk lists. Specifically, it examines whether leveraging these curated fact-based resources enhances the detection accuracy and reliability of LLMs compared to a zero-shot learning approach. The following research questions were designed to address this objective:

RQ1: To what extent does augmenting LLMs with debunk lists improve their performance in detecting health misinformation?

RQ2: How accurately can LLMs apply the use of debunk lists in their reasoning justification when identifying health misinformation?

RQ3: How do LLMs apply debunk entities during misinformation detection across different debunk list sizes?

2. Literature Review

The use of LLMs for detecting misinformation has been increasing in recent years. Researchers have explored the use of LLMs for detecting misinformation from different perspectives including prompting settings [19,34,35], leveraging their multilingual abilities [20,36] and multimodal capabilities, [19,37,38], fine-tuning [37,39], and Retrieval-Augmented Generation (RAG) approaches [21,22]. Perline et al. [20] evaluated the performance of GPT-4 using a zero-shot prompting approach for detecting four-way misinformation classification in a multilingual setting (English and German). GPT-4, under zero-shot settings, obtained better results than those of other benchmarks, reaching an F1 score of 42.8% (English) and 38.7% (German). In another study, Chen et al. [35] used NoCoT and zero-shot CoT strategies on GPT-3.5, GPT-4, and Llama2-7b/13b-chat to detect both human-generated misinformation and LLM-generated misinformation. Their results show that the CoT prompting technique outperforms NoCoT in most of their experiments. Wu et al. [19] leveraged GPT-3.5 as a feature extractor to improve the detection of out-of-context image-caption pairs. Their results show that their proposed method, GPT+RF, outperforms those in previous studies, reaching an accuracy of 92.9%.

Fine-tuning LLMs have also been explored in various studies to improve misinformation detection [31,39]. Pavlyshenko [39] explored fine-tuning Llama 2 for various

disinformation tasks, including fake news detection, propaganda analysis, fact checking, and sentiment extraction. Their results highlight that the fine-tuned Llama-2 model is capable of analyzing complex styles and narratives as well as detecting sentiment that can be used as a predictive feature in machine learning models. In another study, various LLMs, including GPT, Llama, and Gemini, were explored for detecting fake news detection [31] on an ISOT dataset [40]. Their results show that fine-tuned models, such as (RoBERTa with an F1 score of 0.99), significantly outperform LLM models (with an Avg. F1 score of 55.29%). Even fine-tuned Llama 3.1-8B reaches an F1 score of 60%. LoRA, introduced in [41], has been studied as an efficient approach for fine-tuning LLMs in misinformation detection tasks [37,39]. Mura et al. [37] reported that the combination of the LoRa fine-tuning and data augmentation techniques significantly improved the model's performance.

Some studies have explored improving LLMs' detection performance using the RAG method, which incorporates external knowledge and tools [22,42]. RAG can potentially improve fake news detection by combining a retrieval system with an LLM. For instance, Wan et al. (2024) [21] prompted LLMs to identify the key elements in a news article and then retrieve related Wikipedia content. These retrieved passes are then prepended to the original article to improve the LLM's contextual understanding in misinformation detection task. Cheung and Lam's [22] experiments highlight that integrating external knowledge with LLMs can lead to better fact-checking performance. Their proposed method, Fact Lama+, augmented with external knowledge, outperforms previous approaches by reaching an F1 score of 30.44% on the LIAR dataset [43].

The debunk list approach used in this study, while conceptually related to RAG, differs in various important ways. RAG retrieves large volumes of uncured documents based on similarity search. The retrieved content is then fed to a downstream model for further processing [44]. However, this approach does not validate the factual correctness of the retrieved content. Consequently, irrelevant similarities might be included, potentially introducing noise that can negatively impact model performance [45,46]. In contrast, our curated debunk lists were collected from health authorities and contain pre-verified false claims. This provides high-quality and reliable external knowledge that aligns semantically with common health misinformation patterns. Debunk lists allow the model to conceptually match claims rather than rely on surface similarity, like in RAG. This helps reduce hallucination and improve misinformation detection. Although debunk lists require some level of human effort to collect and verify, they are lightweight and fast to implement. New entities can be added immediately without building retrieval infrastructure, making debunk lists especially practical for rapid deployment during emerging health misinformation events. Although RAG is generally more scalable across different domains, debunk lists potentially offer higher accuracy and lower noise.

3. Materials and Methods

This study utilized a multistage methodological framework to evaluate the performance of LLMs in detecting health misinformation, as illustrated in Figure 1. In the first stage (A), a debunk list is created. This list consists of health-related claims that have been verified as false information [47]. The list was extracted from trustworthy health sources such as the World Health Organization (WHO) [48], Centers for Disease Control and Prevention (CDC) [49], and John Hopkins University (JHU) [50]. In the second stage (B), the test data are collected from social media posts. Both datasets are processed through an extract-transfer-load (ETL) process, through which raw data are cleaned, standardized, and formatted for the next phase of prompt design. The transformed data are then stored in a database to be used in experimental settings. In the third stage (C), prompts are designed for the Debunk-Augmented setting, which uses both the test data and the debunk entries.

In the fourth stage (D), prompts include only the test data to represent the baseline zero-shot settings. In the fifth stage (E), prompts are sent to LLMs to detect misinformation on social media posts. In the sixth and final stage (F), the outputs are analyzed and evaluated using standard performance metrics.

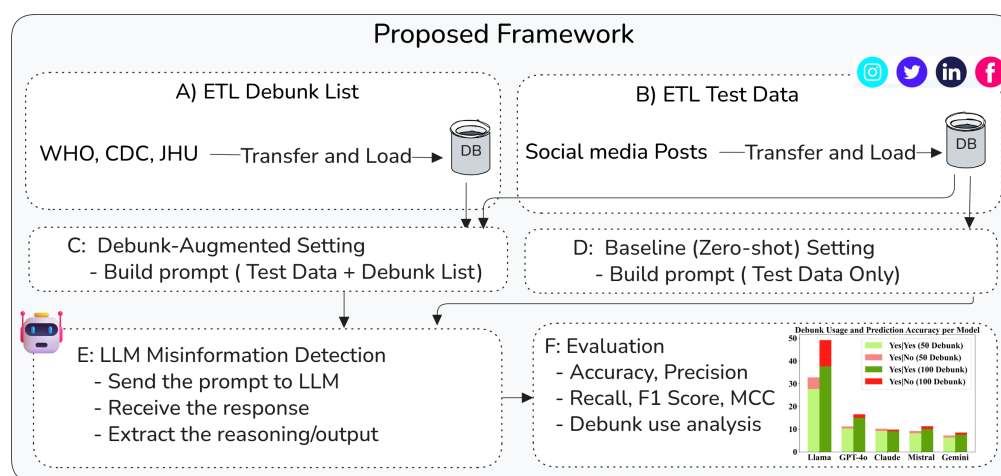


Figure 1. This figure illustrates the proposed framework for evaluating LLMs in detecting health misinformation. In stages (A,B), the debunk list and test data sets are collected, cleaned, and normalized, respectively. The framework includes three experimental settings: stage (C) represents a Debunk-Augmented setting (50 and 100 debunks), where prompts include both test data and debunk lists; stage (D) represents a baseline zero-shot setting that uses only test data. In stage (E), prompts are sent to LLMs, and in stage (F), the responses are analyzed and evaluated.

3.1. Debunk-Augmented Setting

In the debunk setting, prompts include both the statement and debunk items. Models are guided to classify the claims using the debunk list if relevant. If the models cannot find a relevant match, they are instructed to use their general medical knowledge. Output is returned in JSON format, comprising the predicted label, whether a debunk list was used, the ID of the debunk list if applicable, and a brief explanation. A sample prompt is shown in Figure 2.

Classify the following statement referencing the debunk list. If you can not find a relevant debunking entity, use your general medical knowledge to make your decision.

DebunkList: {debunking_list}
Statement: {statement}

You are required to respond in JSON format including the below fields:

- 1) Classification_Label: True or False
- 2) DebunkList_Used: Yes or No
- 3) DebunkList_Num: Debunk Number or 'N/A'
- 4) Justification: In less than 50 words explain your decision based on using debunk list or your general medical knowledge.

Figure 2. This figure illustrates an example prompt used for the Debunk-Augmented setting to guide the LLM to classify a health-related statement using a debunk list or general medical knowledge.

3.2. Zero-Shot Setting

In the zero-shot setting, the prompts are designed to use only the test statements without having access to additional external knowledge or the debunking entities. The models are guided to classify each statement as True or False based on their pretrained knowledge and reasoning. The output is required to be in JSON format, comprising the predicted label and a short explanation. An example prompt is shown in Figure 3.

Classify the following statement applying your medical knowledge, reasoning ability and false claims patterns.
Statement: {statement}

You are required to respond in JSON format including the below fields:
1) Classification_Label: True or False
2) Justification: In less than 50 words explain your decision based on using your general medical knowledge.

Figure 3. This figure illustrates an example prompt used for the zero-shot setting to guide the LLM to classify a health-related statement only using general medical knowledge, without having access to the external debunk information.

3.3. LLM Selection and Setup

In this study, we selected five state-of-the-art LLMs to evaluate their performance in detecting health misinformation. The chosen models were Llama-3.1-8B-instruct [18], Mistral-large-2411 [51], GPT-4o-mini [16], Claude-3.5-haiku-20241022 [52], and Gemini-1.5-flash [17]. Llama-3.1-8B is an 8 billion-parameter model, designed for multilingual dialog and code tasks [53]. Mistral-Large-2411 has 123 billion parameters and is optimized for reasoning, coding, and long-context use [54]. The parameter sizes of Claude-3.5-haiku, Gemini-1.5-flash, and GPT-4o-mini are not publicly disclosed. Claude-3.5-haiku is designed for tasks such as code completions, interactive chatbots, and data extraction and labeling in domains like finance, healthcare, and research [52]. Gemini-1.5-flash is a multimodal model optimized for efficient performance across text, image, audio, video, and long-context inputs [17]. The domain focus for GPT-4o mini has not been publicly specified.

We selected these models because they were released in 2024 and support large-context windows of more than 100k tokens. Large-context windows are critical in this study because they allow the models to process both the statement and debunk list entities in a single prompt. In addition, these models are widely used in the literature, are publicly available, and represent a diverse range of capabilities. This selection enables us to evaluate the generalizability of our framework across different model types while considering computational limitations that affect reproducibility for other researchers. We did not include larger models such as GPT-4o and Llama-3.1-70B due to their high computational cost. The outputs of all models were returned in JSON format through API requests, except for Llama, which was run on a local server.

3.4. Evaluation Setup

The evaluation framework for this study includes assessing the model performance using five standard metrics: accuracy (1), precision (2), recall (3), F1 score (4), and Matthews Correlation Coefficient (MCC) (5) [55,56]. These metrics provide a complete view of model performance, beyond simple correctness. Precision and recall capture the balance between false positives and false negatives, while the F1 score reflects the overall classification quality. MCC evaluates the level of agreement beyond chance.

The performance of each model is evaluated under three experimental conditions: the Debunk-Augmented setting (50 and 100 Debunk) and zero-shot setting. These configurations allow us to measure the impact of external knowledge on misinformation detection. In addition to models' classification performance, we evaluate how well each model recognizes and reports the use of the debunk entities. Specifically, we explore if the models reporting use of debunk information are more likely to generate accurate predictions. Each metric's formula is provided below [55,56]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \quad (5)$$

3.5. Dataset

In this study, we used a publicly available COVID-19 misinformation dataset [57] to test our proposed framework. This dataset includes 10,700 English social media claims about COVID-19, each labeled as either “real” or “fake”. Fake or misleading claims were collected from public fact-checking websites such as PolitiFact, and Snopes, and were manually verified against the original source documents. Real news was collected from verified official twitter accounts, (e.g., ICMR and CDC) and labeled as real if the content included factual information about COVID-19. The dataset is balanced, with 47.66% fake posts and 52.34% real posts. Following standard practice in machine learning, we randomly sampled 20% of the dataset (2,140 claims) to create a balanced, fixed test dataset. This test dataset was sufficiently large and balanced to provide reliable estimation of model’s performance.

In addition, we manually collected 120 debunked false claims about COVID-19 from trustworthy health sources, such as the WHO, CDC, and John Hopkins University, to sample the debunk settings (50 and 100 debunks). Claims were included if they were explicitly labeled as false at least with one of these agencies. Each claim was screened for clarity, and any duplicates were removed. The remaining claims were reformatted into the standard labeling format to maintain consistency across the list. The entities cover the major thematic categories of conspiracy theories, prevention, and treatment. Table 1 summarizes the distribution of the datasets.

Table 1. Distribution of test dataset and debunk list.

Dataset	False	True	Total
Constraint ([57])	5100	5600	10,700
Test dataset (20% of Constraint-COVID-19)	1020	1120	2140
Debunk List (CDC, WHO, JHU)	120	0	120

4. Results

This study evaluated the performance of five state-of-the-art LLMs under three experimental settings: zero-shot and Debunk-Augmented with 50 and 100 items. The goal of these evaluations is to assess the effectiveness of structured external knowledge in improving models’ misinformation detection performance. Model performance is examined across the three configurations to highlight the impact of debunk lists on various classification metrics (RQ1). The analysis also explores whether models correctly identify and apply relevant debunk information in their reasoning justification (RQ2). In addition, the most used debunk list entries are identified to provide insights for developing functional debunk lists for real-world deployment (RQ3).

4.1. Performance Comparison Across Debunk List Size (RQ1)

Table 2 illustrates the performance of five LLMs across five evaluation metrics, including accuracy, precision, recall, F1 score, and MCC, under zero-shot and 50 and 100 Debunk-Augmented items. Overall, the results show that leveraging LLMs with structured external knowledge in the form of debunk lists improves the performance of most models in detecting misinformation. In addition, increasing the number of debunk items from 50 to 100 generally improves the performance further; however, the magnitude of the improvements depends on the model.

Table 2. Performance of LLMs under zero-shot and Debunk-Augmented conditions.

Model	Approach	Accuracy	Precision	Recall	F1 Score	MCC
Claude	Zero-shot	73.30%	66.20%	89.50%	76.10%	50.10%
	50-Debunk	81.60%	78.30%	84.70%	81.40%	63.40%
	100-Debunk	84.96%	83.15%	85.95%	84.47%	70.02%
Gemini	Zero-shot	74.60%	69.00%	84.20%	75.90%	50.60%
	50-Debunk	81.50%	78.10%	84.80%	81.30%	63.20%
	100-Debunk	83.49%	81.68%	84.33%	82.93%	67.07%
GPT-4o	Zero-shot	80.30%	79.50%	78.90%	79.20%	60.50%
	50-Debunk	83.00%	85.50%	77.20%	81.10%	66.00%
	100-Debunk	83.71%	89.39%	74.58%	81.28%	67.96%
Llama	Zero-shot	81.10%	81.20%	78.20%	79.70%	62.00%
	50-Debunk	81.90%	86.10%	73.90%	79.50%	64.10%
	100-Debunk	80.10%	78.16%	82.14%	79.80%	60.79%
Mistral	Zero-shot	84.20%	82.10%	85.40%	83.70%	68.40%
	50-Debunk	86.10%	88.20%	81.60%	84.80%	72.20%
	100-Debunk	88.08%	93.33%	80.70%	86.54%	76.61%

As shown in Figure 4, the most notable improvement is observed in Claude, where the F1 score increased by 6.96% under the zero-shot setting (76.1%) compared to under 50-Debunk (81.4%) and further to 84.5% in 100-Debunk. Similarly, Gemini and Mistral gain a 9.26% and 3.39% improvement in F1 score, respectively, under the zero-shot to 100-Debunk settings. These results highlight the role of a larger debunk size in providing broader misinformation coverage, assisting these models in more accurate classifications. GPT-4o shows lower improvement in F1 score (79.2% to 81.3%), but gains considerable improvement in precision from 79.5% to 89.4% in the 100-Debunk setting. This suggests that the model becomes more cautious in classifying misinformation when exposed to larger external knowledge. In contrast, Llama shows inconsistent results, dropping its F1 score under 50-Debunk (79.7% to 79.5%) and then recovering slightly with 100-Debunk (79.8%). This suggests that the model's design and reasoning process play a critical role in effectively using debunk augmentation.

In addition to the standard evaluation metrics, MCC provides an overall performance of model predictions by considering all four elements of the confusion metrics, as shown in Figure 5. For instance, while GPT-4o shows a slight improvement in F1 score, its MCC increases from 60.5% to 67.96%. This highlights the model's capability for more balanced predictions. Similarly, Mistral reaches the highest MCC of 76.61% in the 100-Debunk setting, which emphasizes not only its high precision and recall but also its consistent predictions over the whole dataset.

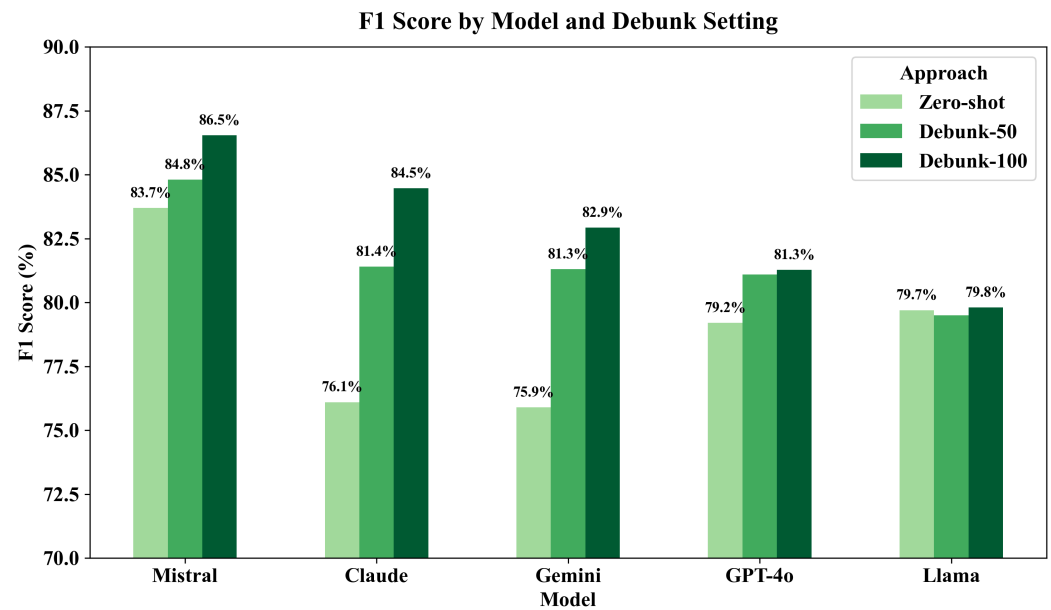


Figure 4. This figure illustrates the F1 score of 5 LLMs across three evaluation settings: zero-shot, Debunk-Augmented (50 items), and Debunk-Augmented (100 items). This chart shows the overall performance improvement for most models when augmented with external knowledge (debunk lists).

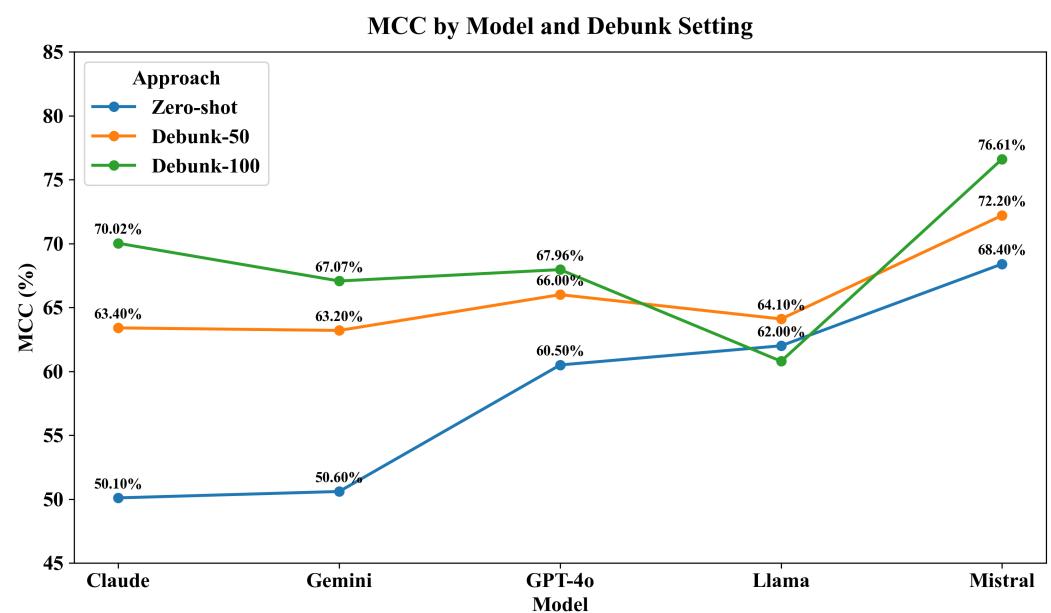


Figure 5. This figure illustrates MCC across five LLMs under the zero-shot, Debunk-50, and Debunk-100 approaches. Higher values indicate stronger overall classification balance.

While augmenting LLMs with debunk lists improves the model's overall performance, it tends to lower the recall for most models. As shown in Table 2 and Figure 6, Claude, GPT-4o, and Mistral decline in recall when augmented with debunk lists. For instance, Claude's recall drops from 89.5% under the zero-shot setting to 84.7% in Debunk-50 and recovers slightly to 85.95% with Debunk-100. GPT-4o shows slower decreases (from 78.9% to 74.6%), while Mistral's recall drops from 85.4% to 80.7%. This suggests that when models are exposed to external knowledge, they may miss more true positives. In contrast, Llama shows an unpredictable pattern by dropping its recall to 73.9% for 50-Debunk but significantly recovering in 100-Debunk.

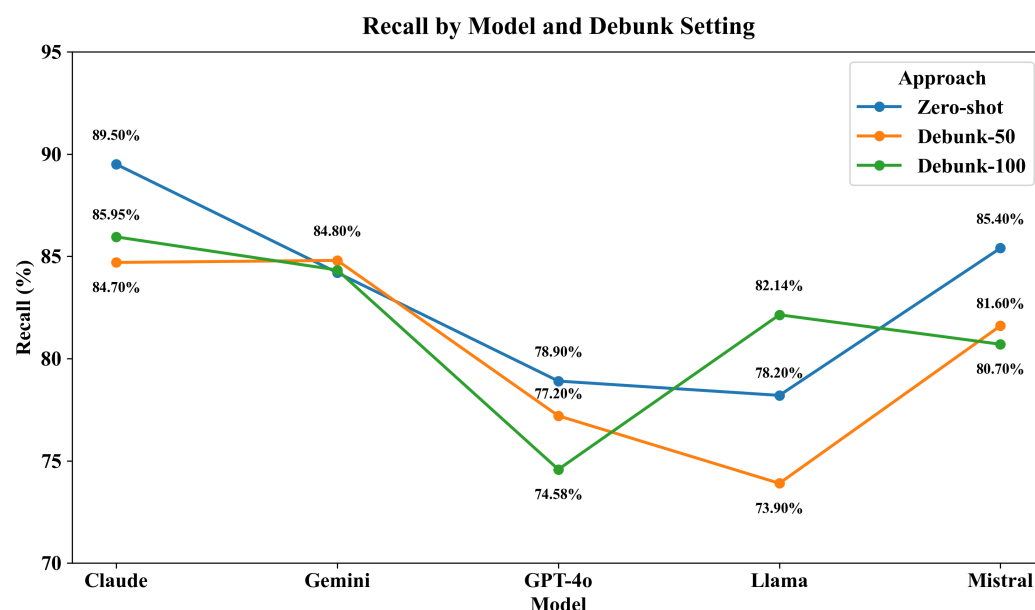


Figure 6. This figure shows the recall performance of five LLMs under zero-shot, Debunk-50, and Debunk-100 settings. Higher recall reflects better coverage of true positives.

Figure 7 shows two sample responses from Claude-3.5-haiku, highlighting how external knowledge can guide the model's predictions, especially in scenarios where the model lacks domain-specific knowledge. The Debunk-Augmented output references a relevant debunk entity (# 15), while the zero-shot output is based only on the model's internal knowledge. This shows the model's strong reasoning and proper use of related debunk items.

Statement: " If you get coronavirus from Chinese food the simple cure is to gargle bleach"
Actual Label: False

Model Output:

```
{ "Classification_Label": "False",
  "Justification": " The statement is false. It perpetuates misinformation by linking COVID-19 to Chinese food and suggesting a dangerous, unproven 'cure.' Gargling bleach is harmful and not a recognized treatment for any illness. Public health guidelines emphasize proper hygiene and medical care for COVID-19."}
```

Statement: " If you get coronavirus from Chinese food the simple cure is to gargle bleach"
Actual Label: False

Debunk_List # 15: You can protect yourself from COVID-19 by injecting, swallowing, bathing in or rubbing onto your body bleach, disinfectants or rubbing alcohols.

Model Output:

```
{ "Classification_Label": "False",
  "DebunkList_Used": "Yes",
  "DebunkList_Num": 15,
  "Justification": "The debunk list entry Num: 15 directly addresses the false claim that bleach can protect against or cure COVID-19. Gargling bleach is extremely dangerous and can cause severe harm."}
```

Figure 7. This figure represents sample responses from Claude under zero-shot (**top**) and Debunk-Augmented (**bottom**) settings for the same test statement. The Debunk-Augmented output references a relevant debunk entity, while zero-shot relies only on the model's internal knowledge.

4.2. Evaluating Debunk Utilization and Correct Prediction Rates (RQ2)

To provide a better understanding of how models apply external knowledge during misinformation detection, we explore the frequency and correct use of debunk list across both 50-item and 100-item debunk list. Table 3 summarizes the percentage distribution for four model outcomes: (1) No | No: no debunk list was used and the model made incorrect predictions; (2) No | Yes: no debunk list was used and the model made correct predictions; (3) Yes | No: a debunk list was used and the model made incorrect predictions; (4) Yes | Yes: a debunk list was used and the model made correct predictions. We focused on the last two categories to investigate if models used debunk lists correctly.

Table 3. This table illustrates the distribution of entities where models use or do not use debunk list (for both 50 items and 100 items) and whether their prediction was correct. In the column labels, the first value (Yes/No) indicates whether the debunk list used, and the second value (Yes/No) indicates whether the prediction was correct. The four columns, therefore, represent the following: no debunk used and incorrect prediction (No | No), no debunk used and correct prediction (No | Yes), debunk used and incorrect prediction (Yes | No), and debunk used and correct prediction (Yes | Yes).

Model	Type	No No	No Yes	Yes No	Yes Yes
Claude	50-Debunk	17.5%	72.4%	1.0%	9.2%
	100-debunk	14.3%	75.9%	0.7%	9.1%
Gemini	50-Debunk	17.6%	75.2%	0.9%	6.3%
	100-debunk	15.6%	75.9%	0.9%	7.6%
GPT-4o	50-Debunk	17.0%	71.8%	0.9%	10.3%
	100-debunk	15.2%	68.3%	1.6%	15.0%
Llama	50-Debunk	13.2%	53.9%	5.1%	27.7%
	100-debunk	8.3%	42.5%	11.6%	37.6%
Mistral	50-Debunk	12.9%	77.9%	1.0%	8.2%
	100-debunk	10.7%	78.0%	1.2%	10.1%

In most models, increasing the debunk list from 50 to 100 items improved correct use (Yes | Yes). However, the magnitude of this improvement varies by model. GPT-4o's correct use of debunk information increases from 10.3% to 15%, which indicates its capability in applying additional debunk items. However, this also increases the incorrect usage from 0.9% to 1.6%, which highlights the trade-off between higher usage and lower precision.

As shown in Figure 8, Llama shows the highest debunk list usage but also the highest rate of incorrect prediction. It shows the most significant change through the increase in Yes | Yes from 27.7% to 37.6% and significant increase of 5.1% to 11.6% in Yes | No. In contrast, Mistral shows the most stable trend through its increasing Yes | Yes rates from 8.2% to 10.1%, while its Yes | No rates increase moderately from 1% to 1.2%. This highlights the ability of Mistral in its correct use of debunk lists. Claude and Gemini show minor changes, which suggests that their performance saturated regardless of the increase in debunk list length. These findings show that while larger debunk lists can improve model performance, they may introduce noise to some models.

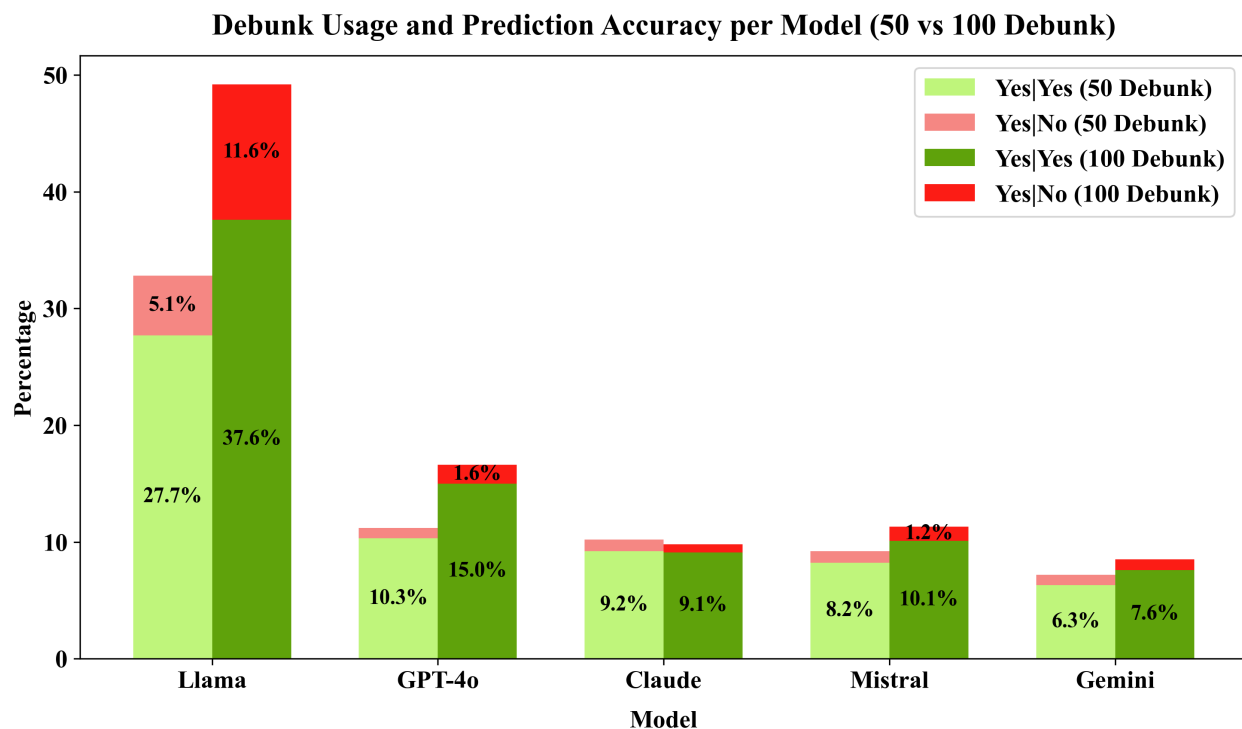


Figure 8. This bar chart shows the proportion of predictions for which a debunk list was used (50 vs. 100 items) and the prediction being either correct (Yes | Yes) or incorrect (Yes | No). It highlights how increasing the size of external knowledge impacts the performance of model prediction.

4.3. Qualitative Analysis of Debunk Usage (RQ3)

To have a better understanding of how LLMs apply the debunk list during misinformation detection, we analyze the top debunk items that were most used during prediction. Our analysis is based on two categories across two debunk list size (50 vs. 100 items): when a debunk item is used and the prediction is correct (Yes | Yes) and when a debunk item is used but the prediction is incorrect (Yes | No). Tables 4 and 5 summarize the most frequently used debunk items for each category.

Table 4. This table summarizes the most frequently used debunks when models made correct predictions (Yes | Yes) across both 50- and 100-item debunk list configurations.

Debunk Setting	Debunk ID	Debunk Text	Claude	Gemini	GPT-4o	Llama	Mistral	Total
50-Debunk	2	Quercetin can protect you from the coronavirus or treat COVID-19	9	9	21	37	11	87
	3	Zinc can protect you from the coronavirus or treat COVID-19		11	11	47		69
	12	Hydroxychloroquine or chloroquine can treat COVID-19	14	14	12	17	14	71
	15	You can protect yourself from COVID-19 by injecting, swallowing, bathing in or rubbing onto your body bleach, disinfectants or rubbing alcohols		7	23	4	30	64
	32	COVID-19 is an engineered biological weapon created by the United States or China	40	16	35	74	36	201
100-Debunk	2	Quercetin can protect you from the coronavirus or treat COVID-19	7	9	16	31	11	74
	3	Zinc can protect you from the coronavirus or treat COVID-19		16	28	70	2	116
	12	Hydroxychloroquine or chloroquine can treat COVID-19	11	12	12	14	11	60
	28	Taking in alcoholic drinks will protect you from getting COVID-19 or treat the disease	31	18	41	83	29	202
	41	Social distancing is not important to reduce the spread of the Covid virus	4	2	4	84	1	95

Table 5. This table summarizes the most frequently used debunks during incorrect prediction (Yes | No) across both the 50- and 100-item settings.

Debunk Setting	Debunk ID	Debunk Text	Claude	Gemini	GPT-4o	Llama	Mistral	Total
50-Debunk	12	Hydroxychloroquine or chloroquine can treat COVID-19	4	3	6	4	5	22
	17	Dexamethasone is a treatment for all COVID-19 patients	1	4		4	3	12
	45	The ingredients in COVID-19 vaccines are not safe	2			14		16
	49	COVID-19 vaccines cause infertility	2	3	1	5	3	14
	50	COVID-19 vaccines alter DNA	1	1		28	2	32
100-Debunk	3	Zinc can protect you from the coronavirus or treat COVID-19			1	27		28
	40	Getting the vaccines to prevent pneumonia and flu prevent COVID-19				63		63
	41	Social distancing is not important to reduce the spread of the Covid virus	1	1		92		94
	45	The ingredients in COVID-19 vaccines are not safe	3	2	3	27	3	38
	46	A negative COVID test means you are safe		1		49		50

4.3.1. Analysis of Debunk Usage in Correct Prediction (Yes | Yes)

As shown in Table 4, specific debunk items were utilized by multiple models during correct predictions. For example, Debunk item #32 (“COVID-19 is an engineered biological weapon created by the United States or China”) and Debunk item #2 (Quercetin can protect you from the coronavirus or treat COVID-19) were used several times across various models. This shows that these debunk entries provide useful semantic cues to the test data and are well aligned with the structure and language of the data. In addition, many of these top-used debunk items reflect the most common concerns raised by the public during a pandemic, such as remedies and the origin of the virus. Moving from 50-Debunk to 100-Debunk results in some interesting shifts. We notice that several top entries from the 50-Debunk setting remain top entries under the 100-Debunk setting (e.g., Debunk items #2, #3, and #12). This indicates that these core entries play a critical role in misinformation detection. However, under the 100-Debunk condition, new top debunk items such as Debunk item #28 (“Taking in alcoholic drinks will protect you from getting COVID-19 or treat the disease”) and Debunk item #41 (“Social distancing is not important to reduce the spread of the Covid virus”) appear. This shift suggests that increasing the number of debunk items to 100 allows the models to have access to broader and more diverse semantic content. This access improves the prediction over a wider range of misinformation topics.

4.3.2. Analysis of Debunk Usage in Incorrect Prediction (Yes | No)

To gain a deeper understanding of models’ misuse of debunk entries, we analyzed which debunk items were used most frequently when the model misclassified the statement (Yes | No). Table 5 summarizes the most frequently misused debunk items across both the 50- and 100-Debunk configurations. Most models, such as Claude, GPT-4o, Gemini, and Mistral, do not show an increased number of incorrect predictions moving from the 50- to 100-Debunk configuration. In fact, expanding the number of debunk entities resulted in lowering the incorrect use of the debunk entities and improving the accuracy of prediction. This indicates that a more diverse debunk sets allows for broader coverage, which helps these models avoid repeating similar mistakes.

However, Llama demonstrates a different pattern. Llama shows a noticeable increase in incorrect prediction when the length of debunk list entries increases from 50 to 100. For instance, Debunk item #41 (“Social distancing is not important to reduce the spread of the Covid virus”) was used 92 times in incorrect predictions. Similarly, Debunk items #40 and #45, which focus on vaccine efficiency, were significantly misused by Llama, which negatively affected its performance. To investigate the root causes of Llama’s misclassifications, we analyzed various model response examples. Table 6 shows a few instances where Llama used the debunk entries that were loosely related to the actual

claims. For example, the input claim “Social distancing results in record low flu numbers” is classified as false, which is justified by Debunk item #41. Another example with factual claims is “The US has now completed tests on over 1 million people”, incorrectly linked to Debunk item #46. Although the claim and debunk entry look similar, Llama misuses the debunk entry. This results in the rejection of a correct statement. These examples suggest that Llama makes errors in the 100-item setting by relying on thematic similarities to incorrectly reject true statements. This behavior of Llama is in contrast with that of the other models, whose accuracy improves by increasing the size of debunk lists.

Table 6. This table provides examples where Llama incorrectly labels true claims as false by incorrectly using the debunk entities.

Claim	Prediction	Justification
Social distancing results in record low flu numbers	False	The statement claims that social distancing results in record low flu numbers. However, debunk list entry 41 states that social distancing is not important to reduce the spread of the Covid virus. This implies that social distancing may not be effective in reducing flu numbers either, making the statement false.
These are the 15 states that reported over 400 new cases	False	The statement implies that social distancing is not important to reduce the spread of the Covid virus, which contradicts general public health guidelines. This is directly related to False Claim #41, which states that social distancing is not important to reduce the spread of the Covid virus.
The US has now completed tests on over 1 million people	False	This is similar to False Claim 46, which states that “A negative COVID test means you are safe,” implying that the number of tests completed is not increasing.

4.4. Comparison with Prior Work

While previous studies have explored strategies to improve misinformation detection using LLMs, direct comparison is challenging due to differences in study design. Table 7 summarizes a high-level overview of recent studies that integrate external knowledge to improve LLM-based misinformation detection. Wan et al. [21] and Li et al. [32] explored misinformation across a range of topics by integrating external evidence from Wikipedia and web domains, respectively. While these studies utilized broad external sources, they did not focus on external health knowledge. Cao et al. [33] focused on scientific misinformation by integrating scientific abstracts from the CORD- database. Although this study focuses on the COVID-19 domain, their work addresses the verification of scientific reports’ accuracy. In contrast, our study focuses on exploring whether augmenting external knowledge with LLMs can improve their ability in detecting health misinformation in social media claims, where data are often noisy, informal, and filled with acronyms.

Unlike previous studies, our research directly evaluates how LLMs can utilize fact-checked false claims, debunk lists, to improve misinformation detection. In addition, we assess not only models’ performance improvement but also whether they can correctly recognize and report their use of external knowledge. This approach offers valuable insights into LLMs’ accuracy, reasoning, and reliability in detecting misinformation.

Table 7. This table compares recent studies using external knowledge to improve LLM-based misinformation detection.

Research Study	External Knowledge	Method	Dataset	F1 Score
Wan et al. [21]	Wikipedia	DELL selective	PHEME	MaF = 0.82
Li et al. [32]	Domain URL	FactAgent with Expert Workflow	PolitiFact	0.88
Cao et al. [33]	CORD_19 Database	SIF	SciNews	0.835
Our Study	CDC, WHO, JHU	Debunk augmentation	Constraint	0.865

5. Discussion

The findings of this study highlight the potential of using debunk lists as structured external knowledge to improve generative AI models, particularly LLMs' performance, in misinformation detection. Interestingly, this improvement is achievable without retraining the model architecture, making it a practical approach for mitigating misinformation that has a dynamic and rapidly evolving nature. By simply augmenting LLMs with debunk lists, the models can be guided to make more accurate classifications, especially in cases where models suffer from a lack of specific domain knowledge.

Increasing the size of a debunk list improved most models' performance; however, the magnitude of improvement varies across models. Claude and Gemini show the highest gains in F1 score. This suggests that these models have stronger reasoning abilities and are better optimized for effectively integrating external knowledge. GPT-4o-mini and Mistral also benefited consistently from the increasing the size of the debunk list but slightly less. On the other hand, Llama shows only minimal improvement between the zero-shot and Debunk-Augmented configurations. Its performance even declined with longer debunk lists. The exact root cause cannot be confirmed without full disclosure of the training details. However, potential explanations include a smaller model size, less training on following instructions through prompts, and sensitivity to longer prompts. These findings suggest that the benefit of debunking augmentation depends on the underlying model structure and training process.

Although most models showed lower recall when augmented with debunk lists, this pattern is expected. The debunk augmentation makes models be more conservative in detecting false claims by prompting them to verify claims against debunk items. As a result, some false claims that are only weakly similar to debunk statements may be missed. This increases false negatives and lowers recall. However, the same reason also reduces false positives and improves the reliability of predictions, as reflected in higher precision and MCC. MCC evaluates classification performances by applying all four elements of a confusion matrix (true positive, true negatives, false positives and false negatives). This allows MCC to capture the overall reliability of the models' prediction when precision and recall move in opposite directions.

A deeper analysis of model justifications shows that models apply debunk lists differently. For instance, while Llama frequently reported the use of debunk lists in its explanations, it also had the highest rate of incorrect predictions. This highlights concerns about the reliability of self-reported usage where higher usage does not necessarily reflect better application. In contrast, other models, such as Claude, Mistral, and Gemini, used the debunk entities less frequently but more effectively. These results show that recognizing external knowledge is not the same as integrating it effectively into reasoning justification.

In addition to performance metrics, the way models relate to debunking entities provides valuable insight into their reasoning behavior. The models' reliance on a small number of effective debunk entries, especially the ones related to common health misinformation topics, like remedies and prevention, highlight the role of specific debunk items as strong guiding cues during decision making. This suggests that a small but well-chosen debunk list can improve model performance without needing a large knowledge base. Creating such a list requires understanding of common misinformation themes on social media. The benefit of debunk lists depends on whether models can apply them correctly. If not, external knowledge may confuse models and reduce performance.

In practice, a debunk list can be integrated into a real-time misinformation monitoring pipeline for social media platforms and online health systems. Debunk lists can be easily updated since new entries can be added quickly during emerging public health events. This allows rapid deployment without the need for a complex retrieval infrastructure. In

addition, this approach can improve existing fact-checking systems by reducing false claims and improving trust in automated detection systems.

6. Limitations and Future Work

While this study shows the effectiveness of debunk list augmentation for detecting health information, it also has some limitations. An additional investigation could examine the inconsistent behavior of Llama-3.1-8B by comparing model size, training methods, and sensitivity to longer prompts. The debunk lists we used in this study mainly focus on COVID-19-related themes, including conspiracies, prevention, treatments, and remedies. These topics reflect the overall public concerns during the early stage of the pandemic. Future work should explore whether expanding debunk lists to cover more diverse misinformation themes, such as political influence or financial health claims, can further improve the generalizability and performance of the models.

While curated debunk lists offer high precision, they require upfront manual effort. Future work should compare RAG implementation and debunk lists from different perspectives such as accuracy, implementation complexity, and risk of introducing noisy content. It would also be beneficial to investigate a hybrid framework that combines the scalability of RAG with the precision of debunk lists.

All LLMs evaluated in this study were released in 2024. They may have been exposed to COVID-19-related content during their pretraining. To reduce this potential effect, we instructed the models in the prompts to first use the debunk list as a reference and only then rely on their general knowledge. Future work could also evaluate models with a pre-COVID-19 knowledge cutoff. This would provide more accurate insights into how much of their performance is related to pretraining exposure versus use of the debunk list. It is also important to highlight that misinformation is shaped by social, cultural, and temporal factors. COVID-19 conspiracy theories spread extensively during the early stage of the pandemic, while the later stages focused more on vaccine and treatment-related content. Although our debunk lists capture some of these shifts, further research should examine how social context, cultural framing, and temporal factors influence the model performance.

Additionally, the proposed framework can be generalized to other misinformation domains, like politics and finance. Future work should explore its adaptability across these areas. However, unlike the health domain, where false claims can be verified from authoritative sources, misinformation in politics or finance is usually more subjective. Applying this framework to other domains may require the development of domain-specific debunk lists from reputable fact-checking organizations. This may also require refining prompts to integrate domain-relevant context.

Author Contributions: Conceptualization, M.R., and S.H.; methodology, M.R. and S.H.; software, M.R.; formal analysis, M.R. and S.H.; data curation, M.R.; writing—original draft preparation, M.R.; writing—review and editing, S.H.; visualization, M.R.; supervision, S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data supporting the findings of this study are publicly available and are cited in the Dataset section of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CDC	Centers for Disease Control and Prevention
ETL	Extract Transfer Load
LLMs	Large Language Models
MCC	Matthews Correlation Coefficient
WHO	World Health Organization

References

1. Van Der Linden, S. Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nat. Med.* **2022**, *28*, 460–467. [CrossRef]
2. Nelson, T.; Kagan, N.; Critchlow, C.; Hillard, A.; Hsu, A. The danger of misinformation in the COVID-19 crisis. *MO Med.* **2020**, *117*, 510–512.
3. Islam, M.S.; Sarkar, T.; Khan, S.H.; Kamal, A.H.M.; Hasan, S.M.; Kabir, A.; Yeasmin, D.; Islam, M.A.; Chowdhury, K.I.A.; Anwar, K.S.; et al. COVID-19-related infodemic and its impact on public health: A global social media analysis. *Am. J. Trop. Med. Hyg.* **2020**, *103*, 1621–1629. [CrossRef]
4. Jennings, W.; Stoker, G.; Bunting, H.; Valgarðsson, V.O.; Gaskell, J.; Devine, D.; McKay, L.; Mills, M.C. Lack of Trust, Conspiracy Beliefs, and Social Media Use Predict COVID-19 Vaccine Hesitancy. *Vaccines* **2021**, *9*, 563. [CrossRef]
5. Su, Z.; McDonnell, D.; Wen, J.; Kozak, M.; Abbas, J.; Šegalo, S.; Li, X.; Ahmad, J.; Cheshmehzangi, A.; Cai, Y.; et al. Mental health consequences of COVID-19 media coverage: The need for effective crisis communication practices. *Glob. Health* **2021**, *17*, 4. [CrossRef] [PubMed]
6. CNN. Israeli Strikes Kill Some of Iran's Most Powerful Men, Including Military and Nuclear Leaders. 2025. Available online: <https://www.cnn.com/2025/06/13/middleeast/israel-iran-strikes-military-deaths-intl-hnk> (accessed on 18 June 2025).
7. IranInternational. Rouhani Claims Islamic Medicine: IRGC Commanders Were Tracked by the Coronavirus Vaccine. 2025. Available online: <https://www.iranintl.com/202506161750?source=share-link> (accessed on 18 June 2025).
8. Ilias, L.; Roussaki, I. Detecting malicious activity in Twitter using deep learning techniques. *Appl. Soft Comput.* **2021**, *107*, 107360. [CrossRef]
9. Khan, T.; Michalas, A. Seeing and Believing: Evaluating the Trustworthiness of Twitter Users. *IEEE Access* **2021**, *9*, 110505–110516. [CrossRef]
10. Theophilo, A.; Giot, R.; Rocha, A. Authorship Attribution of Social Media Messages. *IEEE Trans. Comput. Soc. Syst.* **2023**, *10*, 10–23. [CrossRef]
11. Alghamdi, J.; Lin, Y.; Luo, S. Towards COVID-19 fake news detection using transformer-based models. *Knowl.-Based Syst.* **2023**, *274*, 110642. [CrossRef]
12. Tashtoush, Y.; Al-Rababah, B.; Darwish, O.; Maabreh, M.; Alsaedi, N. A Deep Learning Framework for Detection of COVID-19 Fake News on Social Media Platforms. *Data* **2022**, *7*, 65. [CrossRef]
13. Pangakis, N.; Wolken, S.; Fasching, N. Automated Annotation with Generative AI Requires Validation. *arXiv* **2023**, arXiv:2306.00176. [CrossRef]
14. Imran, M.; Mitra, P.; Srivastava, J. Cross-Language Domain Adaptation for Classifying Crisis-Related Short Messages. *arXiv* **2016**, arXiv:1602.05388.
15. Tan, Z.; Beigi, A.; Wang, S.; Guo, R.; Bhattacharjee, A.; Jiang, B.; Karami, M.; Li, J.; Cheng, L.; Liu, H. Large language models for data annotation: A survey. *arXiv* **2024**, arXiv:2402.13446. [CrossRef]
16. OpenAI. Model. 2025. Available online: <https://platform.openai.com/docs/models> (accessed on 8 October 2025).
17. Gemini. Gemini models. 2025. Available online: <https://ai.google.dev/gemini-api/docs/models> (accessed on 8 October 2025).
18. Meta. Llama 3. 2025. Available online: <https://www.llama.com/models/llama-3/> (accessed on 8 October 2025).
19. Wu, G.; Wu, W.; Liu, X.; Xu, K.; Wan, T.; Wang, W. Cheap-Fake Detection with LLM Using Prompt Engineering. In Proceedings of the 2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Brisbane, Australia, 10–14 July 2023; pp. 105–109. [CrossRef]
20. Pelrine, K.; Imouza, A.; Thibault, C.; Reksoprodjo, M.; Gupta, C.; Christoph, J.; Godbout, J.F.; Rabbany, R. Towards Reliable Misinformation Mitigation: Generalization, Uncertainty, and GPT-4. *arXiv* **2023**, arXiv:2305.14928. [CrossRef]
21. Wan, H.; Feng, S.; Tan, Z.; Wang, H.; Tsvetkov, Y.; Luo, M. Dell: Generating reactions and explanations for llm-based misinformation detection. *arXiv* **2024**, arXiv:2402.10426. [CrossRef]

22. Cheung, T.H.; Lam, K.M. FactLLaMA: Optimizing Instruction-Following Language Models with External Knowledge for Automated Fact-Checking. In Proceedings of the 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Taipei, Taiwan, 31 October–3 November 2023; pp. 846–853. [\[CrossRef\]](#)
23. Rostami, M.; Hossain, K.S.M.T.; Hawamdeh, S. Detecting Health Misinformation by Leveraging LLM Models and Debunk List. In Proceedings of the ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE '25), New York, NY, USA, 24–26 June 2025; pp. 1–6. [\[CrossRef\]](#)
24. Chen, C.; Shu, K. Combating misinformation in the age of llms: Opportunities and challenges. *AI Mag.* **2024**, *45*, 354–368. [\[CrossRef\]](#)
25. Wornow, M.; Lozano, A.; Dash, D.; Jindal, J.; Mahaffey, K.W.; Shah, N.H. Zero-shot clinical trial patient matching with llms. *NEJM AI* **2025**, *2*, A1cs2400360. [\[CrossRef\]](#)
26. Liu, X.; McDuff, D.; Kovacs, G.; Galatzer-Levy, I.; Sunshine, J.; Zhan, J.; Poh, M.Z.; Liao, S.; Achille, P.D.; Patel, S. Large Language Models are Few-Shot Health Learners. *arXiv* **2023**, arXiv:2305.15525. [\[CrossRef\]](#)
27. Wang, Z.; Pang, Y.; Lin, Y. Large Language Models Are Zero-Shot Text Classifiers. *arXiv* **2023**, arXiv:2312.01044. [\[CrossRef\]](#)
28. Parnami, A.; Lee, M. Learning from Few Examples: A Summary of Approaches to Few-Shot Learning. *arXiv* **2022**, arXiv:2203.04291. [\[CrossRef\]](#)
29. Hadi, M.U.; Qureshi, R.; Shah, A.; Irfan, M.; Zafar, A.; Shaikh, M.B.; Akhtar, N.; Wu, J.; Mirjalili, S.; Al-Tashi, Q.; et al. Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Prepr.* **2023**, *1*, 1–26.
30. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* **2025**, *43*, 1–55. [\[CrossRef\]](#)
31. Emil, R.Ş.; BRAD, R. A Comparative Study in Large Language Models Usage for Fake News Detection. *Adv. Artif. Intell. Mach. Learn.* **2024**, *4*, 2810–2823. [\[CrossRef\]](#)
32. Li, X.; Zhang, Y.; Malthouse, E.C. Large language model agent for fake news detection. *arXiv* **2024**, arXiv:2405.01593. [\[CrossRef\]](#)
33. Cao, Y.; Nair, A.M.; Eyimife, E.; Soofi, N.J.; Subbalakshmi, K.; Wullert II, J.R.; Basu, C.; Shallcross, D. Can Large Language Models Detect Misinformation in Scientific News Reporting? *arXiv* **2024**, arXiv:2402.14268. [\[CrossRef\]](#)
34. Pendyala, V.S.; Hall, C.E. Explaining Misinformation Detection Using Large Language Models. *Electronics* **2024**, *13*, 1673. [\[CrossRef\]](#)
35. Chen, C.; Shu, K. Can LLM-Generated Misinformation Be Detected? *arXiv* **2024**, arXiv:2309.13788. [\[CrossRef\]](#)
36. Chalehchaleh, R.; Farahbakhsh, R.; Crespi, N. Enhancing Multilingual Fake News Detection through LLM-Based Data Augmentation. In *Proceedings of the International Conference on Complex Networks and Their Applications, Istanbul, Turkey, 10–12 December 2024*; Springer: Cham, Switzerland, 2024; pp. 258–270. [\[CrossRef\]](#)
37. Mura, D.A.; Usai, M.; Loddo, A.; Sanguinetti, M.; Zedda, L.; Di Ruberto, C.; Atzori, M. Is it fake or not? A comprehensive approach for multimodal fake news detection. *Online Soc. Networks Media* **2025**, *47*, 100314. [\[CrossRef\]](#)
38. Zhong, W.; Xiao, Y.; Xu, M.; Cheng, X. VMID: A Multimodal Fusion LLM Framework for Detecting and Identifying Misinformation of Short Videos. *arXiv* **2024**, arXiv:2411.10032. [\[CrossRef\]](#)
39. Pavlyshenko, B.M. Analysis of Disinformation and Fake News Detection Using Fine-Tuned Large Language Model. *arXiv* **2023**, arXiv:2309.04704. [\[CrossRef\]](#)
40. Ahmed, H.; Traore, I.; Saad, S. Detecting opinion spams and fake news using text classification. *Secur. Priv.* **2018**, *1*, e9. [\[CrossRef\]](#)
41. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685.
42. Kuntur, S.; Wróblewska, A.; Paprzycki, M.; Ganzha, M. Under the Influence: A Survey of Large Language Models in Fake News Detection. *IEEE Trans. Artif. Intell.* **2025**, *6*, 458–476. [\[CrossRef\]](#)
43. Wang, W.Y. “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. *arXiv* **2017**, arXiv:1705.00648. [\[CrossRef\]](#)
44. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: Vancouver, Canada 2020; Volume 33, pp. 9459–9474.
45. Wang, H.; Prasad, A.; Stengel-Eskin, E.; Bansal, M. Retrieval-Augmented Generation with Conflicting Evidence. *arXiv* **2025**, arXiv:2504.13079. [\[CrossRef\]](#)
46. Shen, X.; Billoshmi, R.; Zhu, D.; Pei, J.; Zhang, W. Assessing “Implicit” Retrieval Robustness of Large Language Models. *arXiv* **2024**, arXiv:2406.18134. [\[CrossRef\]](#)
47. iSSUU. Debunking Disinformation. 2025. Available online: https://issuu.com/unicri/docs/handbook_to_combat_cbrn_disinformation/s/17891193 (accessed on 8 October 2025).

48. WHO. Infodemic. 2025. Available online: https://www.who.int/health-topics/infodemic#tab=tab_1 (accessed on 8 October 2025).
49. CDC. COVID-19. 2025. Available online: https://www.cdc.gov/covid/vaccines/myths-facts.html?CDC_AAref_Val=https://www.cdc.gov/coronavirus/2019-ncov/vaccines/facts.html (accessed on 8 October 2025).
50. Hopkins, J. Health. 2025. Available online: <https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus> (accessed on 8 October 2025).
51. Mistral. Frontier AI. In Your Hands. 2025. Available online: <https://mistral.ai/> (accessed on 8 October 2025).
52. Anthropic. Build with Claude. 2025. Available online: <https://docs.anthropic.com/en/docs/about-claude/models/all-models> (accessed on 8 October 2025).
53. Huggingface. Llama-3.1-8B-Instruct. 2024. Available online: <https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct> (accessed on 18 September 2025).
54. Huggingface. Mistral-Large-Instruct-2411. 2024. Available online: <https://huggingface.co/mistralai/Mistral-Large-Instruct-2411> (accessed on 18 September 2025).
55. Chicco, D.; Warrens, M.J.; Jurman, G. The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen’s Kappa and Brier Score in Binary Classification Assessment. *IEEE Access* **2021**, *9*, 78368–78381. [[CrossRef](#)]
56. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [[CrossRef](#)]
57. Patwa, P.; Sharma, S.; Pykl, S.; Guptha, V.; Kumari, G.; Akhtar, M.S.; Ekbal, A.; Das, A.; Chakraborty, T. Fighting an infodemic: COVID-19 fake news dataset. In *Proceedings of the Combating Online Hostile Posts in Regional Languages During Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, 8 February 2021*; Revised Selected Papers 1; Springer: Cham, Switzerland, 2021; pp. 21–29. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.