# Fake News in Sheep's Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks

Jiaying Wu
National University of Singapore
Singapore
jiayingwu@u.nus.edu

Jiafeng Guo
University of Chinese Academy of Sciences
Institute of Computing Technology, CAS
Beijing, China
guojiafeng@ict.ac.cn

Bryan Hooi
National University of Singapore
Singapore
bhooi@comp.nus.edu.sg

## ABSTRACT

It is commonly perceived that fake news and real news exhibit distinct writing styles, such as the use of sensationalist versus objective language. However, we emphasize that style-related features can also be exploited for *style-based attacks*. Notably, the advent of powerful Large Language Models (LLMs) has empowered malicious actors to mimic the style of trustworthy news sources, doing so swiftly, cost-effectively, and at scale. Our analysis reveals that LLM-camouflaged fake news content significantly undermines the effectiveness of state-of-the-art text-based detectors (up to 38% decrease in F1 Score), implying a severe vulnerability to stylistic variations. To address this, we introduce SheepDog, a style-robust fake news detector that prioritizes content over style in determining news veracity. SheepDog achieves this resilience through (1) *LLM-empowered news reframings* that inject style diversity into the training process by customizing articles to match different styles; (2) a *style-agnostic training* scheme that ensures consistent veracity predictions across style-diverse reframings; and (3) *content-focused veracity attributions* that distill content-centric guidelines from LLMs for debunking fake news, offering supplementary cues and potential intepretability that assist veracity prediction. Extensive experiments on three real-world benchmarks demonstrate SheepDog's style robustness and adaptability to various backbones. [1]

## CCS CONCEPTS

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

## KEYWORDS

Fake News; Large Language Models; Adversarial Robustness

**ACM Reference Format:**
Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake News in Sheep's Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671977

[1]Data and code are available at: https://github.com/jiayingwu19/SheepDog.



*KDD '24, August 25–29, 2024, Barcelona, Spain*

ACM ISBN 979-8-4007-0490-1/24/08.
https://doi.org/10.1145/3637528.3671977

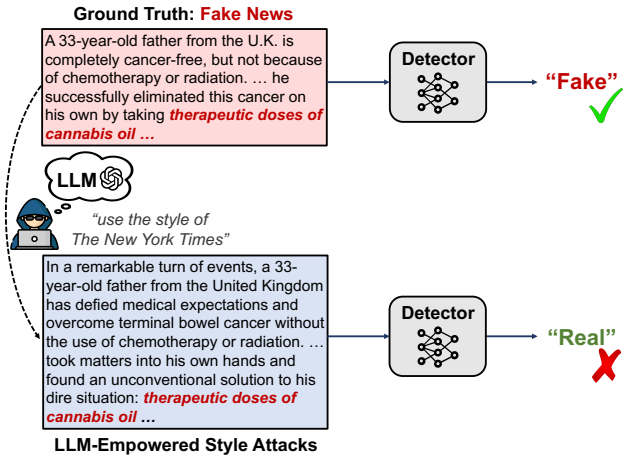


**Figure 1: A motivating example of LLM-empowered style attacks on text-based fake news detectors, where fake news is camouflaged with the style of reliable news publishers.**

## 1 INTRODUCTION

Psychological theories, such as the Undeutsch hypothesis [3], suggest that genuine and fake statements exhibit distinct linguistic styles. Indeed, reputable news sources uphold journalistic integrity, emphasize accuracy and fact-checking, and maintain a balanced tone [5]. In contrast, unreliable outlets often resort to sensationalism, lack credible sources, and may exhibit partisan biases [19]. Building upon these stylistic differences, recent advances in automated fake news detection have incorporated sentiment features [1, 64] to enhance the detector, and highlighted the significance of styles in discerning between hyperpartisan news and well-balanced mainstream reporting [44].

While style-related features serve as key indicators in identifying fake news, they also offer a direct avenue for malicious users to conduct **style-based attacks**. This problem is exacerbated by the advent of powerful Large Language Models (LLMs) [7, 37, 39], whose unprecedented capabilities for reasoning and generative tasks [55, 66] bridges the gap between machine-generated and human-written news. Consequently, malicious actors now possess the capability to mimic the style of reputable news sources, in an attempt to evade automated detection. As shown in Figure 1, using

a style-oriented prompt (i.e., "style of The New York Times"), LLM-camouflaged fake news successfully bypasses a RoBERTa [30]-based fake news detector.

The impact of stylistic variations on fake news detectors has received limited attention, despite prior investigations into attacks regarding social engagements [54], word-level perturbations [27], and machine-generated malicious comments [29]. To assess the robustness of text-based fake news detectors, we introduce a series of style-based attacks, specifically by tailoring news articles to adversarial writing styles (detailed in Section 4.1). Our experiments, as summarized in Table 1, reveal a significant performance degradation in state-of-the-art text-based detectors, with some suffering up to a 38% decline in F1 Score. Existing detectors, well-fitted to real news from trustworthy sources and fake news from unreliable sources, struggle to adapt to real-world scenarios where news content is presented in diverse styles.

To address the *style-related vulnerability* of text-based detectors, we introduce SheepDog, a style-robust approach that consistently recognizes trustworthy content and identifies deceptive content, even when concealed within the LLM-empowered "sheep's clothing". Built upon a pretrained language model (LM) backbone, which can be fully fine-tuned to capture task-specific salient features, and leveraging the strong zero-shot reasoning and generative capabilities of LLMs, SheepDog effectively combines the strengths of both (fine-tuned LM: specialized; LLM: versatile).

The style-agnostic nature of SheepDog stems from its utilization of *LLM-empowered news reframings*. To accommodate the wide spectrum of writing styles presented in real-world news articles, we harness the impressive capabilities of LLMs in adhering to complex real-world instructions [17]. By doing so, we can reframe training articles into style-diverse expressions while preserving the integrity of their content. Subsequently, through *style-agnostic training*, we aim to ensure consistency in the model's veracity predictions across each news article and its style-diverse reframings. This training scheme encourages SheepDog to discount style-related features, enabling it to focus on capturing style-agnostic veracity signals from the news content.

To reinforce the emphasis of our approach on news content over style, we propose to incorporate *content-focused veracity attributions* from an LLM to inform veracity predictions. Leveraging the extensive world knowledge and reasoning capabilities within LLMs [4, 33], we elicit explanatory outputs from LLMs regarding the veracity of news articles, with reference to a set of content-centric fake news debunking rationales specifically related to news content (detailed in Section 5.3). By converting these rationales into precise pseudo-labels, we introduce additional weak supervision that steers SheepDog towards learning robust, style-agnostic news representations. Our approach leverages these attribution-level predictions not only to facilitate style robustness but also to potentially offer explainability into the veracity of news articles.

Our key contributions are summarized as follows:

- **Empirical Finding**: We present a novel finding on the *style-related vulnerability* of state-of-the-art text-based fake news detectors to LLM-empowered style attacks.
- **LLM-Empowered Style Robustness**: We introduce SheepDog, a style-agnostic fake news detector that achieves robustness through style-agnostic training and content-focused veracity attribution prediction, synergized within a multi-task learning paradigm.
- **Effectiveness**: Extensive experiments demonstrate that SheepDog achieves significantly superior style robustness across multiple style-based adversarial settings, and yields consistent performance gains when combined with representative LM and LLM backbones.

## 2 RELATED WORK

**Fake News Detection.** Automated fake news detection has been explored using a wide range of neural architectures [42, 49, 67]. Apart from extracting lexical [45] and sentiment features [44] within the news article text, many methods incorporate auxiliary features to supplement veracity prediction, including user comments [49], news environments [48], knowledge bases [10, 13], temporal patterns from users [46], and social graphs [36, 59, 60]. Recent studies also seek to address challenges including temporal shift [22], entity bias [69] and domain shift [34, 35, 70] in fake news detection scenarios. In this work, we adopt a text-based perspective, specifically focusing on enhancing the robustness of fake news detectors against stylistic variations.

**Adversarial Attack on Fake News Detectors.** Investigating the vulnerabilities of fake news detectors is central to improving their real-world applicability. Hence, existing efforts [15, 20, 27, 32, 54, 68] have studied the impact of different attacks from multiple aspects, including manipulation of social engagements [32, 54] and user behavior [15], fact distortion [27], subject-object exchange [68], and blocking of data availability [20]. However, the impact of writing styles remains underexplored. To bridge this gap, we investigate the resilience of text-based detectors against LLM-empowered style attacks, and propose a style-agnostic solution.

**LLM Capabilities and Misinformation.** LLMs [37, 38, 52] have demonstrated remarkable reasoning capabilities that even match or surpass human performance in certain scenarios [55, 66]. However, the impressive strengths of LLMs have also attracted increasing attention towards LLM-generated misinformation [28]. Recent investigations have found that LLMs can act as high-quality misinformation generators [25, 31, 41, 62], and that LLM-generated misinformation is generally harder to detect [8, 9]. On a related front, recent work explore the role of LLMs as fact-checkers [14, 40, 65] and fake news detectors [9, 43], and leverage the commonsense reasoning capabilities to elicit supplementary explanations from LLMs [4, 18, 21, 33] that facilitate a wide range of tasks. Although we also instruct an LLM to generate style-related adversarial articles, our goal is to simulate real-world scenarios where news are presented in diverse styles. Additionally, instead of leveraging LLMs to make veracity judgments that distinguish false information from the truth [9, 21, 31, 43], in this work, we investigate the role of LLMs in enhancing the style robustness of text-based fake news detectors, specifically through injecting style-diverse reframings and content-centric cues into the training process.

## 3 PROBLEM DEFINITION

Let $\mathcal{D}$ be a news dataset consisting of $N$ questionable news pieces, denoted as $p_1, p_2, \ldots, p_N$. Among the news pieces, $\mathcal{P}_L \subset \mathcal{D}$ is a

set of labeled news articles. Each news article in $\mathcal{P}_L$ is assigned a ground-truth veracity label $y$. In line with prior work [49, 64, 67], $y$ is a binary label that represents either real news or fake news.

As we focus on style-related issues, we consider a *text-based* setting. Formally, the problem can be defined as follows:

Problem 1 (Text-Based Fake News Detection). *Given a news dataset $\mathcal{D}$ with training labels $\mathcal{Y}_L$, the goal is to predict the veracity labels of unlabeled news pieces $\mathcal{P}_U = \mathcal{D} \setminus \mathcal{P}_L$.*

## 4 LLM-EMPOWERED STYLE ATTACKS

In this section, we establish a series of LLM-empowered style attacks, and conduct preliminary analysis to assess the robustness of state-of-the-art text-based fake news detectors.

### 4.1 Attack Formulation

The impressive capabilities of LLMs [7, 37, 39] enable malicious users to disguise fake news with restyling prompts, resulting in camouflaged articles that closely resemble reliable sources. In this work, we explore a direct form of style-based attack utilizing *news publisher names* (e.g., "CNN"). These names possess distinct styles that can be readily adopted by producers of fake news, making them a likely occurrence in real-world scenarios.

To simulate the adversarial situations where news articles are restyled in relation to various publishers, we manipulate the styles of both trustworthy and unreliable news. Specifically, among the test samples, we utlize an LLM to rephrase real news in the style of tabloids, and fake news in the style of mainstream sources. Our general prompt format is shown as follows:

> Rewrite the following article using the style of [publisher name]: [news article]

Based on publisher popularity, in the place of [publisher name], we select "National Enquirer" to transform real news, and "CNN" to transform fake news. These LLM-restyled test articles are then employed to evaluate the resilience of a detector against style-based attacks, as illustrated in Figure 1.

### 4.2 Style-Related Detector Vulnerability

Automated fake news detection becomes increasingly difficult against LLM-empowered style attacks. In this subsection, we conduct preliminary analysis on real-world new articles to evaluate the influence of writing styles on text-based detectors. Our analysis is based on the FakeNewsNet [50] benchmark (consisting of PolitiFact and GossipCop datasets) and the Labeled Unreliable News (LUN) dataset [45], with dataset descriptions relegated to Section 6.1.1 and Table 2. Specifically, we investigate the following question: ***To what extent can text-based fake news detectors withstand LLM-empowered style attacks?***

In Table 1, we examine 13 representative text-based detectors under both original and adversarial settings (detailed method descriptions are provided in Section 6.1.2). These detectors encompass three categories: ***(1)*** *text-based fake news detectors* with diverse task-specific architectures, including Recurrent Neural Networks (RNNs) [49], Convolutional Neural Networks (CNNs) [67], Graph Neural Networks (GNNs) applied to document graphs [53], and Transformers [64]; ***(2)*** *Fine-tuned LMs* [12, 16, 23, 30, 47, 61] on the fake news detection benchmark datasets; and ***(3)*** *LLMs* [37, 39, 52] with zero-shot prompting. Few-shot and fine-tuned LLM experiments are relegated to Appendix A.

**Table 1: Under LLM-empowered style attacks, existing text-based fake news detectors suffer severe performance deterioration in terms of F1 Score (%). (O: original; A (↓): gap between original unperturbed performance and adversarial performance on the test set formulated in Section 4.1).**

| Method | PolitiFact | | GossipCop | | LUN | |
|---|---|---|---|---|---|---|
| | O | A (↓) | O | A (↓) | O | A (↓) |
| dEFEND\c [49] | 82.59 | 12.15 | 70.74 | 4.34 | 80.92 | 19.16 |
| SAFE\v [67] | 79.85 | 8.74 | 70.64 | 2.93 | 79.46 | 13.12 |
| SentGCN [53] | 80.77 | 13.82 | 69.29 | 5.59 | 79.66 | 16.65 |
| DualEmo [64] | 87.76 | 15.34 | 75.36 | 5.89 | 81.52 | 24.97 |
| BERT [12] | 84.99 | 12.68 | 74.50 | 5.52 | 80.96 | 24.61 |
| RoBERTa [30] | 87.40 | 11.23 | 74.05 | 3.05 | 82.12 | 29.65 |
| DeBERTa [16] | 86.30 | 11.73 | 73.80 | 2.85 | 83.67 | 30.34 |
| UDA [61] | 87.74 | 10.14 | 74.22 | 4.54 | 82.94 | 20.71 |
| PET [47] | 85.51 | 11.02 | 74.63 | 3.08 | 83.66 | 31.08 |
| KPT [23] | 87.70 | 13.26 | 74.23 | 2.63 | 84.06 | 31.83 |
| GPT-3.5 [37] | 69.61 | 27.48 | 56.30 | 16.71 | 79.97 | 20.34 |
| InstructGPT [39] | 64.59 | 20.69 | 50.38 | 9.13 | 68.16 | 11.39 |
| LLaMA2-13B [52] | 63.15 | 29.91 | 53.54 | 27.75 | 70.97 | 38.33 |

We evaluate the robustness of detectors against LLM-empowered style attack based on their performance under the adversarial setting outlined in Section 4.1. Our empirical results in Table 1 and Appendix A yield the following two implications:

Observation 1 (Style-Related Vulnerability of Fake News Detectors). *State-of-the-art text-based fake news detectors are susceptible to LLM-empowered style attacks. This susceptibility results in substantial performance degradation,with an F1 Score decline of up to 38.3% on the adversarial test set.*

Observation 2 (Insufficiency of LLMs as Fake News Detectors). *LLMs, despite their impressive zero-shot capabilities as general-purpose foundation models, exhibit inferior detection performance compared to text-based fake news detectors and pre-trained LMs fine-tuned specifically for fake news detection.*

Our two findings suggest a fundamental limitation of text-based fake news detectors in achieving robust veracity predictions against stylistic variations. Detectors overly influenced by styles struggle to reliably differentiate between real and fake news, and even the powerful LLMs may prove inadequate for the specific demands of fake news detection. In the dynamic digital landscape, the styles of news articles evolve rapidly, while the accessibility for malicious users to manipulate style using LLMs exacerbates these variations. Therefore, for effective deployment, a fake news detector must prioritize the assessment of news content over style. This objective motivates our subsequent innovations toward a style-agnostic fake news detection approach.

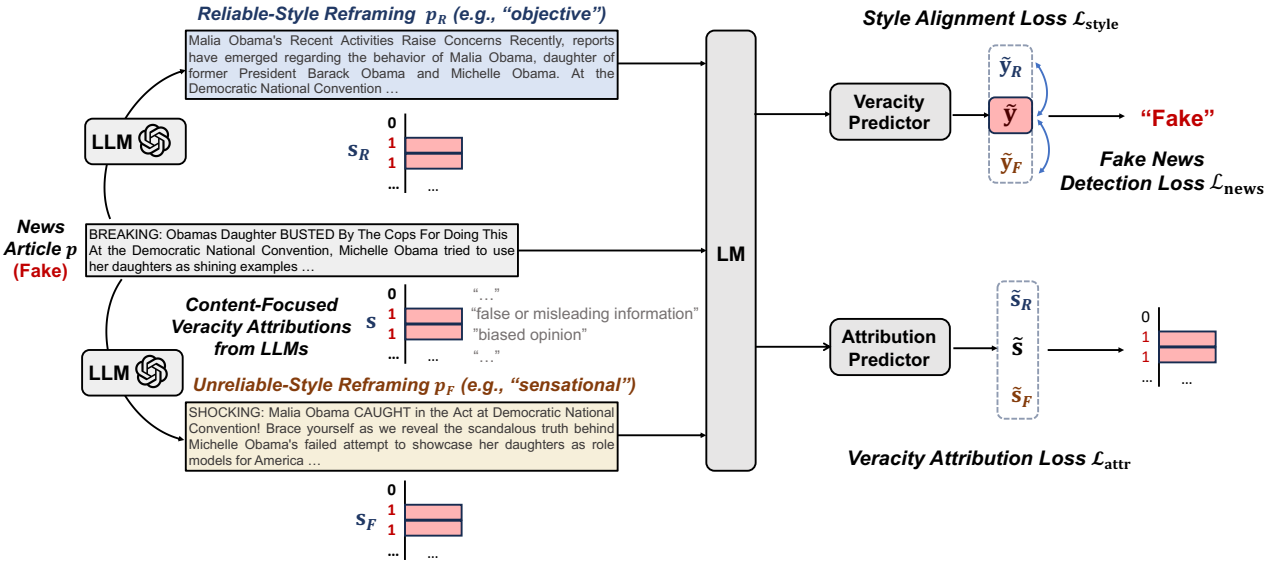


**Figure 2: Overview of the proposed SheepDog framework for style-agnostic fake news detection.**

## 5 PROPOSED APPROACH

Building upon our empirical findings on the style-related vulnerability of text-based fake news detectors, we introduce SheepDog, a style-agnostic detector that reliably assesses news veracity. As overviewed in Figure 2, SheepDog obtains style robustness from two core objectives within a multi-task learning paradigm: ***(1)* style-agnostic training**, which flexibly adapts an LM to the fake news classification task, while ensuring consistent veracity predictions across a diverse array of LLM-empowered news reframings; and ***(2)* content-focused veracity attribution prediction**, which leverages veracity-related insights from LLMs to inform model predictions. To enhance usability and composability, we design our method to be simple and modular, allowing it to be integrated with any LM and LLM backbone.

### 5.1 LLM-Empowered News Reframing

As suggested by our Observation 1, text-based fake news detectors fitted to style-consistent real and fake articles exhibit limited adaptability against stylistic variations. To overcome this limitation, our key idea is to inject style diversity into the training stage through a process we term *reframing*, where each news article is presented in various styles.

SheepDog's reframing strategy is driven by two sub-goals: ***(1)*** encompassing a wide range of styles and ***(2)*** maintaining the integrity of the original news content. LLMs, capable of following complex real-world instructions [17], inherently meet both criteria. This is further validated by our analysis in Appendix C, where LLMs generally prove effective in transforming the tone of news articles while preserving content consistency. Hence, for each training article, we generate a series of prompts, each comprising the news article and a style-oriented reframing instruction. The general structure of the prompt is as follows, with a detailed example presented in Appendix B.4:

> Rewrite the following article in a / an [specified] tone: [news article]

To generate news expressions that simulate both reliable and unreliable sources, we establish a set of four general style-oriented adjectives for the prompt: "objective and professional" and "neutral" to emulate reliable sources, and "emotionally triggering" and "sensational" for unreliable sources. During the training stage, for labeled news article $p \in \mathcal{P}_L$, we randomly select one reliable-style reframing prompt and one unreliable-style reframing prompt to generate diverse expressions. Through querying the LLM, we obtain two corresponding reframings: one reliable-style reframing denoted as $p_R$, and one unreliable-style reframing denoted as $p_F$.

### 5.2 Style-Agnostic Training

A style-robust fake news detector must be capable of discerning the veracity of news articles based on their content, without being influenced by stylistic features. To this end, we introduce a *style alignment objective* that ensures close alignment among the veracity predictions of news article $p$, its reliable-style reframing $p_R$, and its unreliable-style reframing $p_F$. This objective is derived as follows.

Let $\mathcal{M}$ be a pre-trained Language Model (LM) such as RoBERTa [30]. Employing an LM as the backbone of the detector offers advantages, as LMs can be readily fine-tuned for the fake news detection task, which enables them to effectively extract salient task-specific features from the news content. Through $\mathcal{M}$, based on $p$, $p_R$ and $p_F$, we acquire article representations $\mathbf{h} \in \mathbb{R}^d$, and reframing representations $\mathbf{h}_R, \mathbf{h}_F \in \mathbb{R}^d$:

$$\mathbf{h}_p = \mathcal{M}(p), \quad \mathbf{h}_R = \mathcal{M}(p_R), \quad \mathbf{h}_F = \mathcal{M}(p_F). \tag{1}$$

Subsequently, we apply a Multi-Layer Perceptron (MLP) to $p$, $p_R$ and $p_F$ to obtain corresponding veracity predictions $\tilde{\mathbf{y}}, \tilde{\mathbf{y}}_R, \tilde{\mathbf{y}}_F \in \mathbb{R}^2$:

$$\tilde{\mathbf{y}} = \text{MLP}_{pred}(\mathbf{h}), \quad \tilde{\mathbf{y}}_R = \text{MLP}_{pred}(\mathbf{h}_R), \quad \tilde{\mathbf{y}}_F = \text{MLP}_{pred}(\mathbf{h}_F). \tag{2}$$

Each of $\tilde{\mathbf{y}}, \tilde{\mathbf{y}}_R, \tilde{\mathbf{y}}_F$ contain two logits that correspond to the real and fake classes, respectively.

Despite differences in style, the fundamental news content remains consistent across the original news article $p$ and its reframings $p_R$ and $p_F$. Ideally, $p_R$ and $p_F$ should yield the same veracity prediction as $p$. To this end, we formulate the following **style alignment loss** defined as:

$$\mathcal{L}_{\text{style}} = \text{MEAN}\left(\mathcal{L}_1(\tilde{\mathbf{y}}_R, \tilde{\mathbf{y}}),\ \mathcal{L}_1(\tilde{\mathbf{y}}_F, \tilde{\mathbf{y}})\right), \tag{3}$$

where $\mathcal{L}_1$ represents the The Kullback-Leibler (KL) divergence loss.

While aligning the predictions of $p_R$ and $p_F$ with $p$, it is crucial to ensure accurate veracity prediction for $p$. Therefore, we also incorporate a **fake news detection loss**:

$$\mathcal{L}_{\text{news}} = \mathcal{L}_2(\tilde{\mathbf{y}}, y), \tag{4}$$

where $\mathcal{L}_2$ represents the standard cross entropy (CE) loss.

## 5.3 Content-Focused Veracity Attributions

In addition to tuning the LM backbone with the style alignment objective, which discounts style-related features and encourages style-robust predictions, we further propose to integrate auxiliary veracity-related knowledge and reasoning to inform veracity predictions. To achieve this goal, leveraging the impressive zero-shot reasoning capabilities of general-purpose LLMs [4, 33] serves as a promising solution.

Specifically, we elicit *content-focused veracity attributions* from an LLM, which provides explanatory outputs on why each fake news article in the training set is flagged as fake. Our prompt consists of a fake news article and a predefined set $C$ of content-oriented rationales for debunking fake news (e.g., "lack of credible sources" and "false or misleading information"; detailed rationales are described in Appendix B.3). This prompt efficiently leverages the LLM's reasoning capabilities and prior knowledge to identify characteristics associated with fake news, referencing the rationales in $C$. The general prompt format is as follows:

> **Article**: [fake news article]
> **Question**: [given a list of content-centric rationales for debunking fake news, ask the LLM to identify rationales fulfilled by the fake news article]

Upon querying the LLM, we obtain a list of veracity attributions based on the input article. These attributions are then converted into $|C|$-dimensional pseudo-labels, where each rationale in $C$ is represented by a distinct binary label. For a given news article $p \in \mathcal{P}_L$, this process yields pseudo-labels $\mathbf{s} \in \mathbb{R}^{|C|}$, which contains supplementary veracity-related information. Similarly, for reframings $p_R$ and $p_F$, we obtain pseudo-labels $\mathbf{s}_R, \mathbf{s}_F \in \mathbb{R}^{|C|}$, respectively. Notably, since $C$ focuses solely on fake news indicators, the pseudo-labels for real news and its reframings are uniformly set to all zeros.

To distill the veracity-informed knowledge from these attributions, we introduce a *multi-label attribution prediction* objective.

**Table 2: Dataset statistics.**

| Dataset | PolitiFact | GossipCop | LUN |
|---|---|---|---|
| # News Articles | 450 | 7,916 | 7,500 |
| # Real News | 225 | 3,958 | 3,750 |
| # Fake News | 225 | 3,958 | 3,750 |

This enriches our framework with additional content-centric guidance, and offers potential explanability for articles identified as fake during the inference stage (exemplified in Figure 3).

As shown in Eq. 1, we learn news representations $\mathbf{h}, \mathbf{h}_R, \mathbf{h}_F \in \mathbb{R}^d$ for news article $p$ and its reframings $p_R$ and $p_F$, respectively, using a pre-trained LM $\mathcal{M}$. Then, the attribution-level prediction scores $\tilde{\mathbf{s}}, \tilde{\mathbf{s}}_R, \tilde{\mathbf{s}}_F \in \mathbb{R}^{|C|}$ are computed through another MLP:

$$\tilde{\mathbf{s}} = \text{MLP}_{attr}(\mathbf{h}), \quad \tilde{\mathbf{s}}_R = \text{MLP}_{attr}(\mathbf{h}_R), \quad \tilde{\mathbf{s}}_F = \text{MLP}_{attr}(\mathbf{h}_F). \tag{5}$$

The **veracity attribution loss** is then defined as:

$$\mathcal{L}_{\text{attr}} = \text{MEAN}\left(\mathcal{L}_3(\hat{\mathbf{s}},\ \mathbf{s}), \mathcal{L}_3(\hat{\mathbf{s}_R},\ \mathbf{s}_R), \mathcal{L}_3(\hat{\mathbf{s}_F},\ \mathbf{s}_F)\right), \tag{6}$$

where $\mathcal{L}_3$ represents the binary cross entropy (BCE) loss, and $\hat{\mathbf{s}}$ denotes the sigmoid-transformed scores in $\tilde{\mathbf{s}}$ corresponding to each rationale.

## 5.4 Final Objective Function of SheepDog

By enforcing consistency among style-diverse news reframings and exploiting the content-focused attributions from the LLM, the final objective function of SheepDog is defined as a linear combination of the the style alignment loss (Eq. 3), the news classification loss (Eq. 4), and the veracity attribution loss (Eq. 6):

$$\mathcal{L} = \mathcal{L}_{\text{style}} + \mathcal{L}_{\text{news}} + \mathcal{L}_{\text{attr}}. \tag{7}$$

SheepDog is designed as an end-to-end framework, where the style-agnostic news veracity predictor and the content-focused attribution predictor are trained simultaneously.

# 6 EXPERIMENTS

In this section, we empirically evaluate SheepDog to investigate the following six research questions:

- **Robustness Against Style Attacks** (Section 6.2): How robust is SheepDog against LLM-empowered style attacks?
- **Effectiveness on Unperturbed Articles** (Section 6.3): How effectively can SheepDog identify fake news within the original unperturbed test articles?
- **Adaptability to Different Backbones** (Section 6.4): How well does SheepDog perform when combined with different LM and LLM backbones?
- **Ablation Study** (Section 6.5): What are the respective roles of style-agnostic training and content-focused attributions on SheepDog's style robustness?
- **Stability Across Reframing Prompts** (Section 6.6): Does SheepDog yield consistent improvements across diverse sets of news reframing prompts?
- **Case Study** (Section 6.7): How can we interpret SheepDog's rationale for debunking fake news through its predictions on content-focused veracity attributions?

**Table 3: SheepDog significantly outperforms competitive baselines on four adversarial test settings under LLM-empowered style attacks (formulated in Section 4.1), in terms of F1 Score (%) . Bold (underlined) values indicate the best overall (baseline) performance. Statistical significance over the most competitive baselines, computed using the Wilcoxon signed-rank test [56], is indicated with * ($p < .01$). (G1: text-based fake news detectors; G2: LMs fine-tuned to the fake news detection task; G3: LLMs)**

| | Method | PolitiFact | | | | GossipCop | | | | LUN | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A | B | C | D | A | B | C | D | A | B | C | D |
| G1 | dEFEND\c | 70.44 | 69.77 | 73.67 | 72.98 | 66.40 | 66.55 | 68.93 | 69.07 | 61.76 | 62.28 | 72.95 | 72.50 |
| | SAFE\v | 71.11 | 70.80 | 75.55 | 75.24 | 67.71 | 67.05 | 68.31 | 67.65 | 66.34 | 67.08 | 72.40 | 73.16 |
| | SentGCN | 66.95 | 62.50 | 69.54 | 65.08 | 63.70 | 63.07 | 63.61 | 63.01 | 63.01 | 62.50 | 76.11 | 75.56 |
| | DualEmo | 72.42 | 71.23 | 77.07 | 75.80 | 69.47 | 68.50 | 71.69 | 70.71 | 56.55 | 54.78 | 68.53 | 66.80 |
| G2 | BERT | 72.31 | 71.37 | 77.23 | 76.24 | 68.98 | 68.17 | 71.95 | 71.11 | 56.35 | 54.61 | 68.50 | 66.74 |
| | RoBERTa | 76.17 | 74.95 | 78.28 | 77.05 | 71.00 | 70.47 | 72.56 | 72.02 | 52.47 | 53.62 | 68.31 | 69.46 |
| | DeBERTa | 74.57 | 74.36 | 80.60 | 80.35 | 70.95 | 71.15 | 72.51 | 72.71 | 53.33 | 55.45 | 67.16 | 69.27 |
| | UDA | 77.60 | 75.57 | 79.21 | 77.17 | 69.68 | 69.33 | 72.16 | 71.80 | 62.23 | 61.80 | 68.25 | 67.80 |
| | PET | 74.49 | 70.75 | 75.49 | 71.76 | 71.55 | 70.85 | 73.74 | 73.02 | 52.58 | 53.30 | 63.71 | 64.33 |
| | KPT | 74.44 | 73.32 | 77.73 | 76.60 | 71.60 | 71.01 | 73.69 | 73.10 | 52.23 | 53.62 | 65.71 | 67.15 |
| G3 | GPT-3.5 | 42.13 | 43.44 | 56.61 | 58.17 | 39.59 | 38.67 | 48.44 | 47.38 | 59.63 | 61.24 | 65.74 | 67.43 |
| | InstructGPT | 43.90 | 43.90 | 54.21 | 54.21 | 41.25 | 40.18 | 44.26 | 43.12 | 56.77 | 57.15 | 58.93 | 59.32 |
| | LLaMA2-13B | 33.24 | 34.48 | 53.64 | 55.45 | 25.79 | 26.06 | 37.07 | 37.40 | 32.64 | 33.00 | 50.81 | 51.33 |
| **Ours** | SheepDog | **80.99*** | **79.89*** | **82.36*** | **81.24** | **74.45*** | **74.38*** | **75.95*** | **75.88*** | **85.63*** | **86.06*** | **87.89*** | **88.32*** |

**Table 4: Notations and setup for the four style-based adversarial test sets in Section 6.2, denoted as A through D.**

| [publisher name] | CNN | The New York Times |
|---|---|---|
| National Enquirer | A | B |
| The Sun | C | D |

## 6.1 Experimental Setup

*6.1.1 Datasets.* We evaluate our approach on three widely-used real-world benchmark datasets: the FakeNewsNet public benchmark [50], which consists of the **PolitiFact** and **GossipCop** datasets, and the Labeled Unreliable News (**LUN**) dataset [45]. Table 2 describes the dataset statistics. PolitiFact and LUN center on political discourse, while GossipCop focuses on celebrity gossip. For the LUN dataset, which further classifies unreliable news into three sub-categories: satire, hoax, and propaganda, we conduct binary classification between reliable (real) and unreliable (fake) news, and ensure an equal number of unreliable news from each of these fine-grained categories. To better simulate real-world scenarios, we follow prior work [59] and adopt temporal data splitting on PolitiFact and GossipCop, where temporal information is available. The most recent 20% real and fake news articles constitute the test set, and the remaining 80% articles posted earlier serve as the training set. We adopt random 80/20 training / test splits on LUN.

*6.1.2 Baselines.* We benchmark SheepDog against thirteen representative baseline methods, which can be categorized as:

**Text-based fake news detectors (G1)** employ neural architectures tailored specifically for the fake news detection task. **dEFEND\c** is a variant of dEFEND [49] based on the news article text that adopts RNN-based hierarchical co-attention. **SAFE\v** is a text-based variant of SAFE [67] that leverages a CNN-based architecture to learn semantic features. **SentGCN** [53] encodes veracity-related sentence interaction patterns within each article using a GNN, and **DualEmo** [64] incorporates emotion features from news publishers and news comments. As our SheepDog approach does not involve user comments, we implement DualEmo on a BERT-base [12] backbone with publisher emotion features for a fair comparison.

**Fine-tuned LMs (G2)** adapts pre-trained LMs to the fake news detection task, and has proven effective in handling misinformation scenarios [42]. In addition to three widely-recognized LMs, namely **BERT** [12], **RoBERTa** [30], and **DeBERTa** [16], we include **UDA** [61], a representative BERT-based model that employs diverse text augmentations to yield consistent model predictions against input noise. We also select two methods under the popular *prompting* paradigm: **PET** [47], which converts textual inputs into cloze questions that contain a task description; and **KPT**, [24] which expands the label word space with varied class-related tokens. For a fair comparison, as our proposed approach does not involve unlabeled articles, we implement UDA using consistency training on the supervised training data, and exclude the self-training and PLM ensemble components for PET. All methods in this category are implemented with base version LMs, in line with our approach.

**LLMs (G3)** conduct zero-shot veracity prediction. We select three representative baseline LLMs: **GPT-3.5** [37], **InstructGPT** [39], and **LLaMA2-13B** [52], detailed in Appendix B.1.

*6.1.3 Implementation Details.* We implement SheepDog and its variants based on PyTorch 1.10.0 with CUDA 11.1. We utilize pre-trained RoBERTa-base weights from HuggingFace Transformers 4.13.0 [58]. The LM backbone for SheepDog was configured with a maximum sequence length of 512, a batch size of 4, and a learning rate of $2 \times 10^{-5}$. We prompt GPT-3.5 to generate news reframings,

content-focused veracity attributions, and adversarial test articles. (detailed prompting descriptions and examples are presented in Appendix B.2 and B.3). For SheepDog's attribution prediction and veracity prediction, we employ two MLPs, each with a single layer (we also implement a variant with 2-layer MLPs in Section 6.5). Our model is fine-tuned for 5 epochs. For the implementation of baseline methods, we adhere to the architectures and hyperparameters recommended by their respective authors.

We evaluate model performance using Accuracy (%) and macro-F1 Score (%). For all experiments except those involving LLMs, we report averaged metrics over 10 runs of each method to provide a comprehensive evaluation. In the case of LLM zero-shot predictions, we employ greedy decoding and conduct each experiment once.

## 6.2 Robustness Against Style Attacks

We establish a series of LLM-empowered style attacks to assess SheepDog's robustness. Following our prompt template formulated in Section 4.1, in the place of [publisher name], we select "National Enquirer" and "The Sun" to camouflage real news, and "CNN" and "The New York Times" for fake news, according to publisher popularity. This yields $2 \times 2 = 4$ distinct adversarial test sets, labeled as A through D in Table 4. ***Note that we report the results of SheepDog on adversarial set A in all other subsections, unless otherwise specified.***

Table 3 compares the performance of SheepDog with competitive baselines across adversarial test sets A through D under LLM-empowered style attacks. We can observe that: ***(1)*** All baseline methods are highly susceptible to LLM-empowered style attacks. This vulnerability suggests that existing methods exhibit a tendency towards over-fitting on style-related attributes. ***(2)*** UDA, which leverages back-translation to generate diverse text augmentations, consistently demonstrates higher robustness compared to its BERT backbone. This suggests the efficacy of incorporating augmentations. However, UDA still struggles to fully adapt to significant stylistic variances in the input articles. This limitation may be attributed to the fact that augmentations through back-translation alone cannot provide sufficient variance. ***(3)*** On the challenging adversarial test sets of LUN, CNN-based SAFE\v and GNN-based SentGCN are more robust than LM-based baselines, which suggests that LMs can be more prone to overfit to style-related features. ***(4)*** SheepDog outperforms the most competitive baseline by significant margins. Across the three benchmarks, this improvement averages to 2.59%, 2.77%, and 15.70% across the four adversarial test sets, in terms of F1 score. The significantly greater improvements on LUN might be attributed to dataset-specific stylistic attributes, toward which we present a detailed discussion in Section 6.8.

## 6.3 Effectiveness on Unperturbed Articles

A desirable fake news detector should achieve style robustness under adversarial settings without compromising its effectiveness under the unperturbed setting. Our empirical results, presented in Table 5, demonstrate that SheepDog excels in this regard. When tested on the original, unaltered articles, SheepDog consistently matches (on PolitiFact and GossipCop) or surpasses (on LUN) the performance of the most competitive baseline, in terms of both accuracy and F1 score. Similar to our observation ***(4)*** in Section 6.2,

**Table 5: SheepDog achieves performance (%) that is comparable or superior to competitive baselines on the unperturbed original test sets. Bold (underlined) values indicate the best overall (baseline) performance, and $^*$ indicates $p < .01$ using the Wilcoxon signed-rank test [56].**

| Method | PolitiFact | | GossipCop | | LUN | |
|---|---|---|---|---|---|---|
| | Acc. | F1 | Acc. | F1 | Acc. | F1 |
| dEFEND\c | 82.67 | 82.59 | 70.85 | 70.74 | 81.33 | 80.92 |
| SAFE\v | 79.89 | 79.85 | 70.71 | 70.64 | 79.93 | 79.46 |
| SentGCN | 81.11 | 80.77 | 69.38 | 69.29 | 80.07 | 79.66 |
| DualEmo | 87.78 | <u>87.76</u> | <u>75.51</u> | <u>75.36</u> | 81.78 | 81.52 |
| BERT | 85.22 | 84.99 | 74.60 | 74.50 | 81.13 | 80.96 |
| RoBERTa | <u>88.00</u> | 87.40 | 74.14 | 74.05 | 82.53 | 82.12 |
| DeBERTa | 86.33 | 86.30 | 73.86 | 73.80 | 84.01 | 83.67 |
| UDA | 87.77 | 87.74 | 74.28 | 74.22 | 83.02 | 82.94 |
| PET | 85.56 | 85.51 | 74.75 | 74.63 | 84.00 | 83.66 |
| KPT | 87.78 | 87.70 | 74.38 | 74.23 | <u>84.40</u> | <u>84.06</u> |
| GPT-3.5 | 71.11 | 69.61 | 61.49 | 56.30 | 80.67 | 79.97 |
| InstructGPT | 67.78 | 64.59 | 58.33 | 50.38 | 70.87 | 68.16 |
| LLaMA2-13B | 65.56 | 63.15 | 55.74 | 53.54 | 72.47 | 70.97 |
| SheepDog | **88.44** | **88.39** | **75.77** | **75.75** | **93.05**$^*$ | **93.04**$^*$ |

**Table 6: On different LM backbones, SheepDog demonstrates stable and significant improvements (in F1 %). Statistical significance over the respective LM backbone is computed using the Wilcoxon signed-rank test [56], denoted by $^*$ ($p < .01$).**

| Method | PolitiFact | GossipCop | LUN |
|---|---|---|---|
| RoBERTa | 76.17 | 71.00 | 52.47 |
| SheepDog-RoBERTa | **80.99**$^*$ | **74.45**$^*$ | **85.63**$^*$ |
| BERT | 72.31 | 68.98 | 53.97 |
| SheepDog-BERT | **81.37**$^*$ | **73.54**$^*$ | **80.36**$^*$ |
| DeBERTa | 74.57 | 70.95 | 53.33 |
| SheepDog-DeBERTa | **81.10**$^*$ | **73.89**$^*$ | **82.58**$^*$ |

the significant performance gains on LUN might also stem from dataset-specific stylistic features (detailed in Section 6.8).

## 6.4 Adaptability to LM / LLM Backbones

To assess the flexibility of SheepDog, we evaluate the performance of SheepDog combined with three representative LMs: RoBERTa, BERT and DeBERTa. We also evaluate RoBERTa-based SheepDog combined with two representative LLMs: the closed-source GPT-3.5 and the open-source LLaMA2-13B.

As demonstrated in Table 6, SheepDog ***(1)*** substantially enhances the performance of each respective LM backbone. Additionally, as shown in Table 7, SheepDog ***(2)*** also achieves superior style robustness when utilizing both closed-source and open-source LLMs. This adaptability highlights SheepDog's style robustness from style-invariant training and content-focused attribution prediction, implying its practical utility in real-world scenarios where different LMs and LLMs may be preferred or more readily available.

**Table 7: Leveraging closed-source and open-source LLM backbones, SheepDog demonstrates stable and significant improvements (in F1 %). Statistical significance over the fine-tuned RoBERTa backbone is computed using the Wilcoxon signed-rank test [56], denoted by $^*$ ($p < .01$).**

| Method | PolitiFact | GossipCop | LUN |
|---|---|---|---|
| SheepDog | **80.99**$^*$ | **74.45**$^*$ | **85.63**$^*$ |
| SheepDog-LLaMA2 | 80.82$^*$ | 74.04$^*$ | 81.87$^*$ |
| RoBERTa | 76.17 | 71.00 | 52.47 |

**Table 8: Ablation of SheepDog demonstrates benefits of LLM-empowered news reframing (denoted as R) and content-focused veracity attributions (denoted as A) in F1 Score (%).**

| Method | PolitiFact | GossipCop | LUN |
|---|---|---|---|
| SheepDog | **80.99** | **74.45** | **85.63** |
| w/ 2-layer MLP | 79.83 | 74.03 | 84.75 |
| - R | 76.71 | 70.98 | 53.27 |
| - A | 80.73 | 73.74 | 84.83 |
| RoBERTa | 76.17 | 71.00 | 52.47 |

## 6.5 Ablation Study

To gain deeper insights into the functioning of SheepDog and the role of its different components, we compare SheepDog with the following three model variants:

- **SheepDog w/ 2-layer MLP**, which employs 2-layer MLPs with hidden size of 64 as attribution detector and veracity detector.
- **SheepDog-R**, which excludes style-diverse news reframings and the style-agnostic training component.
- **SheepDog-A**, which excludes content-focused veracity attributions and the attribution prediction component.

Results in Table 8 suggest that: ***(1)*** SheepDog-R without news reframings only yields slight improvements over fine-tuned RoBERTa, which suggests the key role of diverse reframings in SheepDog's robustness. ***(2)*** While the improvement of SheepDog over SheepDog-A may seem slight, incorporating veracity attributions assists in guiding the model to prioritize content over style. Furthermore, SheepDog's attribution predictor equips the framework with explanatory outputs during inference stage, facilitating easier human verification in real-world scenarios (see Figure 3 and Section 6.7 for a concrete illustration of this functionality). ***(3)*** SheepDog, utilizing one single layer for veracity prediction and attribution predictions, slightly outperforms the variant employing 2-layer MLPs. This suggests that the expressiveness of LMs, harnessed through our proposed objectives, yields article representations that contain rich indicators related to both attributions and veracity.

## 6.6 Stability Across Reframing Prompts

As detailed in Section 5.1, we utilize an LLM to generate two types of reframings for a given training article $p$: a reliable-style reframing denoted as $p_R$, and an unreliable-style reframing denoted as $p_F$. To assess SheepDog's stability across different reframing prompts, we examine its performance using four diverse combinations of prompts, denoted as R1 through R4.

Recall that we adopt the following template for news reframing:

> Rewrite the following article in a / an [specified] tone: [$p$]

For R1 through R4, the specified tones are defined as: ($p_R$ / $p_F$)

- **R1**: "objective and professional" /"emotionally triggering".
- **R2**: "objective and professional" / "sensational".
- **R3**: "neutral" / "emotionally triggering".
- **R4**: "neutral" / "sensational".

As shown in Table 9, SheepDog consistently demonstrates stable and significant improvements over the most competitive baseline. This validates the generalizability of our approach, suggesting its potential in effectively combating deceptive information in the ever-evolving digital landscape. Notably, our SheepDog approach conducts sampling between two reliable-style reframings and two unreliable-style reframings, leading to stronger versatility.

**Table 9: Across different sets of news reframing prompts, SheepDog demonstrates stable and significant improvements over the most competitive baseline (in F1 %).**

| Method | PolitiFact | GossipCop | LUN |
|---|---|---|---|
| Baseline (Best) | 77.60 | 71.60 | 66.34 |
| SheepDog | **80.99** | **74.45** | 85.63 |
| w/ R1 | 79.02 | 74.22 | 77.42 |
| w/ R2 | 79.93 | 74.16 | **86.18** |
| w/ R3 | 80.36 | 73.55 | 76.77 |
| w/ R4 | 79.71 | 74.01 | 85.55 |

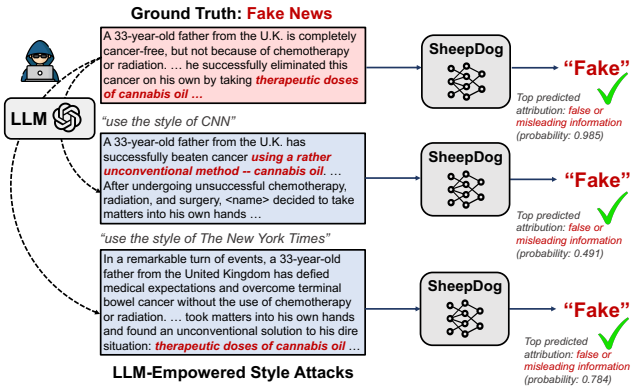


**Figure 3: Across the original fake news article and its LLM-camouflaged counterparts, SheepDog maintains consistency and accuracy in both its veracity prediction and the top-predicted veracity attribution for debunking fake news.**

## 6.7 Case Study

To illustrate SheepDog's potential for offering explanatory outputs during model inference via its attribution predictions (detailed in

Section 5.3), we present a case study based on a fake news article from the LUN test set. As shown in Figure 3, the article falsely claims the effectiveness of cannabis oil in treating cancer, aiming to mislead readers, despite contradicting established medical knowledge. While the baseline RoBERTa detector correctly flags the original article as fake news, it misclassifies two style-transformed adversarial articles as real news. In contrast, SheepDog accurately identifies the original article as fake news, a prediction that remains consistent for its two adversarial counterparts. Remarkably, leveraging the softmax-converted probabilities from attribution-level prediction scores (Eq. 5), SheepDog consistently identifies "false and misleading information" as the top-predicted attribution for debunking fake news. This style robustness is invaluable for practitioners seeking to comprehend the rationale behind each flagged fake news, aiding in human verification and assessment of prediction reliability.

### 6.8 Discussion: Why is SheepDog Yielding Greater Performance Gains on LUN?

SheepDog shows significantly greater performance improvements on LUN compared to PolitiFact and GossipCop in both adversarial (Section 6.2) and original unperturbed settings (Section 6.3). Specifically, it achieves ***[A]*** significant improvements on the *original unperturbed* LUN test set while maintaining performance comparable to the best baseline on PolitiFact and GossipCop (Table 5); and ***[B]*** notably greater improvements on LUN compared to PolitiFact and GossipCop on *style-based adversarial* test sets (Table 3).

These phenomena can be attributed to the unique style-related features of LUN, which include distinct writing styles of ***(1)*** individual *news publishers* and and ***(2)*** different *news types*. (e.g., hoax). Unlike PolitiFact and GossipCop, LUN's publishers (i.e., news sites) in the training and test sets do not overlap [45], creating an inherent distribution shift between training and test data. Furthermore, as described in Section 6.1.1, the fake news articles in LUN encompass satire, hoax, and propaganda with distinctive writing styles. As a result, fake news detectors trained on LUN are expected to be more reliant on writing style.

Recall that our Observation 1 reveals the heavy reliance of existing text-based detectors on styles rather than news content for veracity prediction. Trained on LUN, the baseline models become overly reliant on styles specific to both publishers and news types, which potentially explains [A] and [B]:

- On the *original unperturbed LUN test set* (Table 5), publisher-specific style features used by baseline models fail to generalize to test articles, as test articles are produced by news sites not included in the training data. In contrast, SheepDog, being style-agnostic, remains unaffected by changes in news publisher styles and yields significant improvements.
- On the *adversarial LUN test sets* (Table 3), both publisher-specific and news type-specific style features utilized by baselines fail to generalize. Notably, LLM-empowered style attacks reverse the styles of both reliable and unreliable news (Section 4.1). These style variations make type-specific style features detrimental for veracity prediction. In contrast, SheepDog achieves style robustness through LLM-empowered news reframings and content-focused veracity attributions, thereby reliably detecting fake news.

## 7 CONCLUSION AND FUTURE WORK

In this paper, we address the critical aspect of style-related robustness in fake news detection. Motivated by our empirical finding on the susceptibility of state-of-the-art text-based detectors to LLM-empowered style attacks, we introduce SheepDog, a style-agnostic fake news detector that emphasizes content veracity over style. Jointly leveraging the strengths of task-specific LM backbones and versatile general-purpose LLMs, SheepDog adopts a multi-task learning paradigm, which integrates style-agnostic training and content-focused veracity attribution prediction. Extensive experiments on three real-world benchmarks demonstrate SheepDog's robustness and effectiveness across various style-based adversarial settings, news reframing prompts, and representative backbones. Moving forward, SheepDog lays a solid foundation for developing more resilient and adaptable models in the ever-changing online landscape, and demonstrates promising potential to be further extended to multi-modal scenarios.

## 8 ACKNOWLEDGEMENTS

This work was supported in part by the National Research Foundation Singapore, NCS Pte. Ltd. and National University of Singapore under the NUS-NCS Joint Laboratory (Grant A-0008542-00-00).

## REFERENCES

[1] Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. 2019. Sentiment Aware Fake News Detection on Online Social Networks. In *ICASSP*. 2507–2511.
[2] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31, 2 (2017), 211–36.
[3] Bárbara G. Amado, Ramón Arce, and Francisca Fariña. 2015. Undeutsch hypothesis and Criteria Based Content Analysis: A meta-analytic review. *The European Journal of Psychology Applied to Legal Context* 7, 1 (2015), 3–12.
[4] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *ICLR*.
[5] Philipp Bachmann, Mark Eisenegger, and Diana Ingenhoff. 2021. Defining and Measuring News Media Quality: Comparing the Content Perspective and the Audience Perspective. *The International Journal of Press/Politics* 27 (2021).
[6] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*. 632–642.
[7] Tom Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language Models are Few-Shot Learners. In *NeurIPS*, Vol. 33. 1877–1901.
[8] Canyu Chen and Kai Shu. 2023. Combating Misinformation in the Age of LLMs: Opportunities and Challenges. *arXiv preprint arXiv: 2311.05656* (2023).
[9] Canyu Chen and Kai Shu. 2024. Can LLM-Generated Misinformation Be Detected?. In *ICLR*.
[10] Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation. In *KDD*. 492–502.
[11] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In *NeurIPS*.
[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. 4171–4186.
[13] Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2021. KAN: Knowledge-aware Attention Network for Fake News Detection. *AAAI* 35, 1 (2021), 81–89.
[14] Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2023. Language Models Hallucinate, but May Excel at Fact Verification. arXiv:2310.14564 [cs.CL]
[15] Bing He, Mustaque Ahamad, and Srijan Kumar. 2021. PETGEN: Personalized Text Generation Attack on Deep Sequence Embedding-Based Classification Models. In *KDD*. 575–584.
[16] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *ICLR*.
[17] Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Lida Chen, Xintao Wang, Yuncheng Huang, Haoning Ye, Zihan Li, Shisong Chen, Yikai Zhang, Zhouhong Gu, Jiaqing Liang, and Yanghua Xiao. 2024. Can

Large Language Models Understand Real-World Complex Instructions?. In *AAAI*. 18188–18196.
[18] Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. 2023. Harnessing Explanations: LLM-to-LM Interpreter for Enhanced Text-Attributed Graph Representation Learning. arXiv:2305.19523 [cs.LG]
[19] Nolan Higdon. 2020. What is Fake News? A Foundational Question for Developing Effective Critical News Literacy Education. *Democratic Communiqué* 279 (2020). Issue 1.
[20] Benjamin D. Horne, Jeppe Nørregaard, and Sibel Adali. 2019. Robust Fake News Detection Over Time and Attack. *ACM Trans. Intell. Syst. Technol.* 11, 1, Article 7 (2019).
[21] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. In *AAAI*. 22105–22113.
[22] Beizhe Hu, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Zhengjia Wang, and Zhiwei Jin. 2023. Learn over Past, Evolve for Future: Forecasting Temporal Trends for Fake News Detection. In *ACL*. 116–125.
[23] Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. 2022. Knowledgeable Prompt-tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification. In *ACL*. 2225–2240.
[24] Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. 2022. Knowledgeable Prompt-tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification. In *ACL*. 2225–2240.
[25] Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023. Faking Fake News for Real Fake News Detection: Propaganda-Loaded Training Data Generation. In *ACL*. 14571–14589.
[26] Anna-Katharina Jung, Björn Ross, and Stefan Stieglitz. 2020. Caution: Rumors ahead—A case study on the debunking of false information on Twitter. *Big Data & Society* 7 (2020).
[27] Camille Koenders, Johannes Filla, Nicolai Schneider, and Vinicius Woloszyn. 2021. How Vulnerable Are Automatic Fake News Detection Methods to Adversarial Attacks? arXiv:2107.07970 [cs.CL]
[28] Sarah Kreps, R. Miles McCain, and Miles Brundage. 2022. All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science* 9, 1 (2022), 104–117.
[29] Thai Le, Suhang Wang, and Dongwon Lee. 2020. MALCOM: Generating Malicious Comments to Attack Neural Fake News Detection Models. In *ICDM*. 282–291.
[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL]
[31] Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting Fire with Fire: The Dual Role of LLMs in Crafting and Detecting Elusive Disinformation. In *EMNLP*. 14279–14305.
[32] Yuefei Lyu, Xiaoyu Yang, Jiaxin Liu, Sihong Xie, Philip Yu, and Xi Zhang. 2023. Interpretable and Effective Reinforcement Learning for Attacking against Graph-based Rumor Detection. In *IJCNN*. 1–9.
[33] Sachit Menon and Carl Vondrick. 2023. Visual Classification via Description from Large Language Models. In *ICLR*.
[34] Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. MDFEND: Multi-domain fake news detection. In *CIKM*. 3343–3347.
[35] Qiong Nan, Danding Wang, Yongchun Zhu, Qiang Sheng, Yuhui Shi, Juan Cao, and Jintao Li. 2022. Improving Fake News Detection of Influential Domain via Domain- and Instance-Level Transfer. In *COLING*. 2834–2848.
[36] Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. FANG: Leveraging Social Context for Fake News Detection Using Graph Representation. In *CIKM*. 1165–1174.
[37] OpenAI. 2022. ChatGPT: Optimizing language models for dialogue.
[38] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
[39] Long Ouyang, Jeffrey Wu, Xu Jiang, et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*, Vol. 35. 27730–27744.
[40] Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-Checking Complex Claims with Program-Guided Reasoning. (2023), 6981–7004.
[41] Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the Risk of Misinformation Pollution with Large Language Models. In *Findings of EMNLP 2023*. 1389–1403.
[42] Kellin Pelrine, Jacob Danovitch, and Reihaneh Rabbany. 2021. The Surprising Performance of Simple Baselines for Misinformation Detection. In *WWW*. 3432–3441.
[43] Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. 2023. Towards Reliable Misinformation Mitigation: Generalization, Uncertainty, and GPT-4. arXiv:2305.14928 [cs.CL]
[44] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *ACL*. 231–240.
[45] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *EMNLP*. 2931–2937.
[46] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A Hybrid Deep Model for Fake News Detection. In *CIKM*. 797–806.
[47] Timo Schick and Hinrich Schütze. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *EACL*. 255–269.
[48] Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom Out and Observe: News Environment Perception for Fake News Detection. In *ACL*. 4543–4556.
[49] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. DEFEND: Explainable Fake News Detection. In *KDD*. 395–405.
[50] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data* 8, 3 (2020), 171–188.
[51] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.* 19, 1 (2017), 22–36.
[52] Hugo Touvron, Louis Martin, Kevin Stone, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]
[53] Vaibhav Vaibhav, Raghuram Mandyam, and Eduard Hovy. 2019. Do Sentence Interactions Matter? Leveraging Sentence Level Representations for Fake News Classification. In *TextGraphs-13*. 134–139.
[54] Haoran Wang, Yingtong Dou, Canyu Chen, Lichao Sun, Philip S. Yu, and Kai Shu. 2023. Attacking Fake News Detectors via Manipulating News Social Engagement. In *WWW*. 3978–3986.
[55] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* (2022).
[56] Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1 (1945), 80–83.
[57] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *NAACL*. 1112–1122.
[58] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-Art Natural Language Processing. In *EMNLP*. 38–45.
[59] Jiaying Wu and Bryan Hooi. 2023. DECOR: Degree-Corrected Social Graph Refinement for Fake News Detection. In *KDD*. 2582–2593.
[60] Jiaying Wu, Shen Li, Ailin Deng, Miao Xiong, and Bryan Hooi. 2023. Prompt-and-Align: Prompt-Based Social Alignment for Few-Shot Fake News Detection. In *CIKM*. 2726–2736.
[61] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised Data Augmentation for Consistency Training. In *NeurIPS*, Vol. 33. 6256–6268.
[62] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending Against Neural Fake News. In *NeurIPS*, Vol. 32.
[63] Amy X. Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B. Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, Ed Bice, Sandro Hawke, David Karger, and An Xiao Mina. 2018. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. In *Companion Proceedings of WWW*. 603–612.
[64] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining Dual Emotion for Fake News Detection. In *WWW*. 3465–3476.
[65] Xuan Zhang and Wei Gao. 2023. Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method. arXiv:2310.00305 [cs.CL]
[66] Haoyi Zheng and Huichun Zhan. 2023. ChatGPT in Scientific Writing: A Cautionary Tale. *The American Journal of Medicine* 136, 8 (2023), 725–726.e6.
[67] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-Aware Multi-modal Fake News Detection. In *PAKDD*. 354–367.
[68] Zhixuan Zhou, Huankang Guan, Meghana Bhat, and Justin Hsu. 2019. Fake News Detection via NLP is Vulnerable to Adversarial Attacks. In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence*.
[69] Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022. Generalizing to the Future: Mitigating Entity Bias in Fake News Detection. In *SIGIR*.
[70] Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen Zhuang. 2022. Memory-Guided Multi-View Multi-Domain Fake News Detection. *IEEE Transactions on Knowledge and Data Engineering* (2022).

# A DISCUSSION: EFFECT OF REFRAMINGS ON LLM STYLE ROBUSTNESS

This section investigates the effects of incorporating our style-diverse news reframings (Section 5.1) on LLM style robustness. Following the experimental setup for evaluating style robustness (Section 6), ***we report results on adversarial set A (formulation described in Table 4) unless otherwise specified***. We explore two methods for adapting LLMs to diverse news styles: ***(1)*** **in-context learning** and ***(2)*** **fine-tuning**, with prompt templates and LLM configurations detailed in Appendix B.1. To mitigate randomness, we present averaged metrics from three runs of each approach.

**In-Context Learning.** Table 10 compares a zero-shot GPT-3.5 detector with the following three in-context learning variants: ***(1)*** **GPT-3.5+ICL-2**, which incorporates 2 randomly selected in-context samples (1 real and 1 fake). ***(2)*** **GPT-3.5+ICL-2-R**, which incorporates the same 2 in-context samples as GPT-3.5+ICL-2, enriched with style-diverse reframings. For each in-context sample, we randomly select one reliable-style reframing and one unreliable-style reframing. ***(3)*** **GPT-3.5+ICL-4**, which incorporates 4 randomly selected in-context samples (2 real and 2 fake).

**Table 10: GPT-3.5 in-context learning performance.**

| Method | PolitiFact | GossipCop | LUN |
|---|---|---|---|
| GPT-3.5 | 42.13 | 39.59 | 59.63 |
| + ICL-2 | 38.57 | 38.88 | 57.51 |
| + ICL-2-R | 41.76 | **41.83** | **60.29** |
| + ICL-4 | **42.35** | 41.47 | 60.25 |

Results in Table 10 indicate three key findings: ***(1)*** *LLMs benefit from more in-context samples*, as GPT-3.5+ICL-4 outperforms GPT-3.5+ICL-2. ***(2)*** *Incorporation of news reframings effectively enhances LLM style robustness*, as shown by the consistent gains of GPT-3.5+ICL-2-R over GPT-3.5+ICL-2. ***(3)*** *On the fake news detection task, LLM in-context learning does not offer clear benefits over LLM zero-shot.* The long sequence lengths of news articles limit the number of in-context demonstrations, resulting in insufficient context and overly lengthy prompts that LLMs struggle to process.

**Fine-Tuning.** Table 11 compares a zero-shot LLaMA2-13B detector with the following two fine-tuned variants: ***(1)*** **LLaMA2-13B+FT**, which fine-tunes LLaMA2 using the original training articles. ***(2)*** **LLaMA2-13B+FT-R**, which fine-tunes LLaMA2 using the original training articles and their style-diverse reframings. For each article, we randomly incorporate one reliable-style reframing and one unreliable-style reframing as fine-tuning data.

**Table 11: LLaMA2-13B fine-tuning performance.**

| Method | PolitiFact | GossipCop | LUN |
|---|---|---|---|
| LLaMA2-13B | 33.24 | 25.79 | 32.64 |
| + FT | 36.96 | 47.83 | 34.31 |
| + FT-R | **44.44** | **56.06** | **42.22** |

Both LLaMA2 variants are fine-tuned using QLoRA [11] for 1 epoch with a batch size of 16 and a learning rate of $1 \times 10^{-4}$. From Table 11, we observe: ***(1)*** *Fine-tuning generally enhances task-specific LLM capabilities on fake news detection*, as indicated by consistent improvements of LLaMA2-13B+FT over LLaMA2-13B. ***(2)*** *Incorporating style-diverse news reframings as fine-tuning samples further improves style robustness*, as indicated by consistent improvements of LLaMA2-13B+FT-R over LLaMA2-13B+FT. ***(3)*** *LLaMA2-13B+FT-R still falls short compared to GPT-3.5 on the LUN adversarial test set and performs worse than task-specific detectors and fully fine-tuned LMs in Table 3.* This suggests the need for larger, task-specific corpora for fully adapting LLMs to fake news detection, further highlighting our motivation for leveraging LLM general-purpose capabilities in a zero-shot manner, which are reasonably strong and more readily available.

# B LLM PROMPTING CONFIGURATIONS

## B.1 LLM Baselines

In section 6.1.2, we select three representative LLMs as fake news detection baselines: **GPT-3.5** [37] (model name: `gpt-3.5-turbo-0301`), **InstructGPT** [39] (model name: `gpt-3.5-turbo-instruct`), and **LLaMA2-13B** [52] (model name: `Llama-2-13b-chat-hf`).

For GPT-3.5 and InstructGPT, we use their APIs from OpenAI, and set the temperature to 0 for stable veracity predictions. For LLaMA2-13B, we employ model weights from HuggingFace Transformers [58] version 4.31.0, with `do_sample` set to False for greedy decoding. All models adopt the following prompt for zero-shot fake news detection:

> **Question**: Does the following contain real or fake news (or information)? Answer in one word with either 'Real' or 'Fake', then explain why. [news article]
> **Answer**: [starts with a predicted veracity label]

## B.2 Obtaining SheepDog's News Reframings

Recall from Section 5.1 that we generate reliable-style reframings and unreliable-style reframings with GPT-3.5 for each labeled news article. Among the four prompt templates presented below, we use the first two templates for reliable-style reframings, and the other two for unreliable-style reframings.

During reframing generation, we set the temperature to 0.7 and limit the maximum number of response tokens to 512. See Appendix B.4 for a detailed reframing example.

> Rewrite the following article in an objective and professional tone: [news article]

> Rewrite the following article in a neutral tone: [news article]

> Rewrite the following article in an emotionally triggering tone: [news article]

> Rewrite the following article in a sensational tone: [news article]

### B.3 Obtaining SheepDog's Veracity Attributions

In Section 5.3, we use an LLM to elicit auxiliary content-centric information directly related to news veracity. Specifically, we prompt the LLM to provide explanatory outputs for each fake news article in the training set, based on predefined content-focused rationales aimed at debunking fake news.

We draw from existing literature, which commonly emphasizes two key aspects: ***(1)* sources of information** [26, 63], and ***(2)* correctness of news content** [2, 51]. Inspired by these insights, we devise four rationales:

- **Source**-related rationales: ***[A]*** lack of credible sources; and ***[B]*** inconsistencies with reputable sources. These rationales are informed by prior research emphasizing the importance of reputable sources in validating conclusions [63] and correcting false information [26].
- **Content**-related rationales: ***[C]*** false or misleading information; and ***[D]*** biased opinion. Fake news, defined as intentionally and verifiably false information [2, 51], often manipulates biased opinions to exploit cognitive vulnerabilities in news consumers. Consequently, we devise ***[C]*** in alignment with the definition of fake news, and ***[D]*** to represent a common strategy employed by fake news producers.

Our four rationales cover the key aspects of source- and content-related indicators in debunking fake news. We utilize the following prompt to obtain content-focused veracity attributions, and set the temperature to 0 to ensure stability. An example is presented in Appendix B.4.

> **Article**: [fake news article]
> **Question**: Which of the following problems does this article have? Lack of credible sources, False or misleading information, Biased opinion, Inconsistencies with reputable sources. If multiple options apply, provide a comma-separated list ordered from most to least related. Answer "No problems" if none of the options apply.

### B.4 Reframing and Attribution Examples

Due to the camera-ready page limit, detailed illustrations of SheepDog's news reframings and veracity attributions are provided in Table 13 and Table 14 of the arXiv version of our work[2].

## C ANALYSIS ON CONTENT CONSISTENCY OF LLM-EMPOWERED NEWS REFRAMINGS

LLM-empowered news reframings play a key role in achieving style robustness. Appendix B.4 provides a comparison between a news article and its reliable-style reframing, illustrating concretely the impressive capability of LLMs to introduce different tones while preserving the news content.

To offer a more comprehensive assessment of content consistency between the original news article and its reframings, we investigate *claim entailment* for a straightforward and quantifiable estimation. Specifically, the original article should ideally entail the central factual claims within its reframing, and vice versa.

[2]Available at: https://arxiv.org/pdf/2310.10830.

Given that news articles exhibit significantly longer sequences and more complex logical structures compared to the sentence pairs in natural language inference (NLI) benchmarks (e.g., SNLI [6] and MultiNLI [57]), we opt against utilizing NLI models pre-trained on these benchmarks. Instead, *we instruct a GPT-3.5 model to extract claims and infer claim entailment*. The prompt template for claim extraction is as follows:

> Extract and summarize the central factual claim in the following article. Article: [news article A]. Claim:

Querying the LLM yields a succinct summarization of the article's central factual claim, typically composed of several sub-claims. An example response is provided as follows:

> **Response Example for Claim Extraction:** Financial experts are concerned about the negative impact of China's undervalued yuan on both Asia and the United States. They are calling on regional governments and the Group of 20 leaders to take action to prevent potential currency and trade wars. The experts emphasize the need for neighboring countries to urge China to relax its exchange rate controls in order to address the global current account imbalance. They also highlight the adverse effects of US monetary easing and China's low exchange rate on emerging market economies. The experts are urging the G20 to address the currency problem at its upcoming summit in South Korea and to oppose unilateral devaluation moves and support currency stability.

With the extracted claim, we predict claim entailment using the following prompt:

> **Question**: Does the following article entail the claim: [claim extracted from news article A]? Answer in one word with either 'Yes' or 'No'. Article: [news article B].

While precise entailment assessment requires detailed analysis of each article's nuances and specific claim wordings, evaluating the proportion of article pairs with predicted claim entailment provides a general estimation of content consistency. To ensure comprehensive coverage, we analyze two larger-scale benchmark datasets: GossipCop and LUN (dataset statistics in Table 2).

**Table 12: Claim entailment (%) between original news articles and their objective-style reframings.**

| Dataset | original-entail-objective | objective-entail-original |
|---|---|---|
| GossipCop | 86.20 | 89.17 |
| LUN | 89.22 | 87.53 |

As shown in Table 12, the proportions of claim entailment range from 86.20% to 89.22% across both datasets, suggesting a reasonably high consistency between the central factual claims presented in the original news articles and their reframings. This consistency further validates the effectiveness of our reframing approach in preserving the core content of news while injecting stylistic variations.