

Is LLMs Hallucination Usable?

LLM-based Negative Reasoning for Fake News Detection

Chaowei Zhang¹, Zongling Feng¹, Zewei Zhang², Jipeng Qiang¹, Guandong Xu³, Yun Li^{1*}

¹Yangzhou University

²Auburn University

³The Education University of Hong Kong

{cwzhang,jpqiang,liyun}@yzu.edu.cn,

mz120220995@stu.yzu.edu.cn, zzw475622@gmail.com, gdxu@eduhk.hk,

Abstract

The questionable responses caused by knowledge hallucination may lead to LLMs' unstable ability in decision-making. However, it has never been investigated whether the LLMs' hallucination is possibly usable to generate negative reasoning for facilitating the detection of fake news. This study proposes a novel supervised self-reinforced reasoning rectification approach - SR³ that yields both common reasonable reasoning and wrong understandings (negative reasoning) for news via LLMs reflection for semantic consistency learning. Upon that, we construct a negative reasoning-based news learning model called - *NRFE*, which leverages positive or negative news-reasoning pairs for learning the semantic consistency between them. To avoid the impact of label-implicated reasoning, we deploy a student model - *NRFE-D* that only takes news content as input to inspect the performance of our method by distilling the knowledge from *NRFE*. The experimental results verified on three popular fake news datasets demonstrate the superiority of our method compared with three kinds of baselines including prompting on LLMs, fine-tuning on pre-trained SLMs, and other representative fake news detection methods.

Introduction

Recently, Large Language Models (LLMs), such as GPT-4o, Claude 3, and Llama 3.1 represent a significant lead forward in the field of AI, especially for NLP. The LLMs trained on vast amounts of human-generated text can deeply understand and interpret prompts as well as generate comprehensive, coherent, and contextually relevant reasoning for the prompts (Yang et al. 2024; Bang et al. 2023), which allow them to be applicable to various NLP downstream tasks including fake news detection. Specifically, LLMs can provide multi-aspect aids for fake news detection, including contextual analysis, content understanding, fact-checking, source verification, reasoning generation, and others (Liu et al. 2024b; Pelrine et al. 2023). However, some limitations caused by the problem of "knowledge hallucination" (Chen et al. 2023) are still unavoidable while applying LLMs to identify fake news. The up-to-date related studies are

*Yun Li is the corresponding Author.

Fake News: Donald Trump has been disqualified for running for president.

Positive Reasoning: (can **decrease** the credibility of the news)

The claim is **highly unlikely and lacks concrete evidence**. Firstly, there is no official statement or announcement from FEC or any other agency regarding Trump's disqualification. Secondly, Trump's campaign team has **not issued any statements confirming such a development**.

Negative Reasoning: (can **increase** the credibility of the news)

Speculation surrounds Donald Trump's disqualification from the presidential race, suggesting it's a **strategic move to avoid primary defeat**. This exit lets Trump focus on his business without election loss humiliation.

Real News: White House lit in rainbow colors following us supreme court ruling that same-sex couples have right to marry.

Positive Reasoning: (can **increase** the credibility of the news)

This gesture **promotes equality and respect for all, upholds justice and human rights, supports marginalized communities**, and symbolizes hope, acceptance, and progressive values in the ongoing pursuit of a more perfect union.

Negative Reasoning: (can **decrease** the credibility of the news)

The report on the White House's rainbow illumination might be **politically motivated to divert attention from other issues**. Criticism arises due to the **lack of concrete evidence, on credible sources**, raising doubts about authenticity and potential exaggeration of the claim.

Figure 1: The examples of news' positive and negative reasoning requested via LLMs prompting. Current studies are devoted to avoiding the occurrence of negative reasoning, yet we initially propose to leverage the ability of LLM hallucinations to generate negative reasoning coupled with positive reasoning for adversarial semantic consistency learning.

devoted to eliminating hallucinations via RAG (Retrieval-Augmented Generation) (Boumber et al. 2024), knowledge graph (Guo et al. 2023), or other tricks, yet it is a kind of food for thought that **is it possible to use hallucination to do something for news?** For example, can we leverage both news' correct (a.k.a., positive) reasoning and the incorrect (a.k.a., negative) reasoning generated through LLM hallucinations for adversarial semantic consistency learning?

Generally, knowledge hallucination is a response generated by LLM, which contains "false or misleading information presented as fact" (Ji et al. 2023). Classical LLM reasoning usages for the tasks of fake news detection attempt to

generate correct understanding (positive reasoning) as well as alleviate the impact of hallucinations via the tricks mentioned above (Wang et al. 2024). However, it remains controversial that **"Are hallucinations really useless?"** As we know, the AI researchers usually take advantage of data augmentation (Yin et al. 2024; Wu et al. 2024) and data perturbation (Wang et al. 2023; Ma et al. 2022) techniques to "poison" original data to strengthen the quality, and diversity of training datasets for improving the robustness and generalizability of models. Inspired by that, we raise the concern that **"Is it possible to utilize LLMs' knowledge hallucination as negative data samples?"** To verify this possibility, we initially propose to leverage LLM hallucination to generate negative reasoning for news. In this case, it would be beneficial to prevent fake news detection models from experiencing the issue of "Out-Of-Distribution" (Liu et al. 2024a) by including negative reasoning of news. Fig. 1 demonstrates the examples of the generated reasoning via Llama 3. It is noteworthy that the negative reasoning is false information caused by LLM hallucinations.

To explore the ability of LLMs hallucination in generating negative reasoning for assisting fake news detection, we utilize *Llama 3* to generate both positive and negative reasoning for news in a supervised manner. Specifically, qualified positive reasoning can reinforce news' polarity, yet negative reasoning is expected to decrease the confidence level and even alter the polarity of the news. To this end, we present a **Self-Reinforced Reasoning Rectification** approach, called SR^3 , to force LLMs to output high-quality positive and negative reasoning by considering (1) the polarity transformation and (2) the varying confidence score of news. Based on that, we construct a negative reasoning-based news learning model - *NRFE* that maintains two BERT Encoders for learning the semantic consistency between news and reasoning. To avoid the impact of supervised reasoning requests, we design a student model - *NRFE-D* by distilling the knowledge from *NRFE* for prediction. Overall, the contributions achieved by this study are summarized as follows:

- To our best knowledge, this study originally explores the effect of LLM hallucination in generating negative reasoning aiming at fake news detection. Moreover, we propose a novel self-reinforced reasoning rectification approach (SR^3) via LLM reflection to generate qualified positive and negative reasoning for news.
- We deploy a local fake news detection model complying with the design of knowledge distillation to avoid the impact of supervised reasoning requests, which contains a teacher-side module - *NRFE* and a student-side module - *NRFE-D*. Specifically, *NRFE* involves two BERT Encoders to acquire news-reasoning pairs for semantic consistency learning, and *NRFE-D* is a reasoning-free model that distills the representations of news yielded from *NRFE* for prediction.
- The results verified on three fake news datasets demonstrate that our approach considerably outperforms three types of baselines, thereby highlighting its superiority in exploring LLMs' knowledge hallucination for generating negative reasoning.

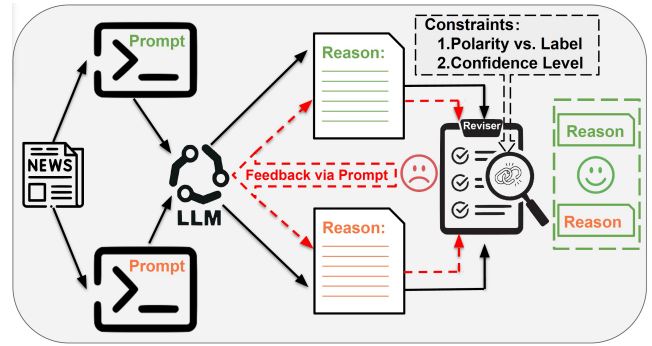


Figure 2: The demonstration of our proposed self-reinforced reasoning rectification approach - SR^3 . Besides positive reasoning requests, SR^3 also utilizes the ability of LLM hallucination to iteratively generate negative reasoning until the generated negative reasoning satisfies the two constraints. More details of SR^3 can be found in Algorithm 1.

Literature Review

Large Language Models (LLMs), such as GPT-4o, offer comprehensive and distinct advantages for fake news detection (Pan et al. 2023a; Chen and Shu 2024). Benefiting from the vast amounts of real-world training data and advanced AI techniques (e.g., Transformer Framework, Reinforcement Learning), LLMs have demonstrated their powerful NLP abilities in perspectives of fact-checking (Pan et al. 2023b), text paraphrasing (Qiang et al. 2023b,a), pattern recognition, and other CoT-recommendable tasks, such as question-answering (Zhang et al. 2024), explanation generation (Bhattacharjee et al. 2024; Wan et al. 2024), reasoning (Jin et al. 2024), etc. One of the up-to-date studies utilizes GPT-3.5 to generate the reasoning with respect to news' common sense and description. (Hu et al. 2024). LLMs also can diversely solo fake news detection via few-shot (Zeng and Gao 2024; Ye and Durrett 2022) or zero-shot (Mitchell et al. 2023) prompting. A recent representative research (Zhang and Gao 2023) aims to mitigate the omission of thoughts and fact hallucination in fake news detection that merely manipulates LLMs for the news' claim verification by taking advantage of in-context learning (ICL), CoT, and external knowledge. In summary, LLMs have manifested their superiority in fake news detection via various prompting tips, yet their powerful usability in fake news detection is still being explored.

Compared with the most recent related studies, this paper proposes a self-reinforced reasoning rectification approach (SR^3) that can iteratively rate the utility of generated positive and negative reasoning by feeding controversies into the next prompting iteration. Fig. 2 exhibits the process of our proposed LLM-based reasoning generation approach for news. It is observable that the SR^3 acts as a reasoning reviser by assessing the polarity (i.e., real or fake) and confidence level (i.e., how much real or fake) of reasoning-based news to verify if and how the generated reasoning can impact the original news.

Data Pre-Processing

Problem Formalization

Suppose a randomly selected news and the corresponding label are denoted by $\langle x, y \rangle$, where $y \in \{Real, Fake\}$. To ensure the effectiveness of generated reasoning in impacting the credibility of news, we first utilize a locally deployed LLM (OLlama 3 70B) to rate the credibility score of x , denoted as $V_{initial}$, where V ranges from 0 (fake) to 100 (real). Then, we produce positive and negative reasoning by feeding LLM with multiple arguments in prompts. Eq.1 illustrates the insights of the reasoning generation.

$$\Psi(x, y, T, V_{initial}, S) = (R, V), \quad (1)$$

$$(R, V) \in \{(R^p, V^p), (R^n, V^n)\}$$

where Ψ embraces a prompt template that takes five arguments as inputs: (1) the news x , (2) corresponding label y , (3) the type of requested reasoning $T \in \{positive, negative\}$, (4) the credibility score of non-reasoning-based news $V_{initial}$, and (5) the alteration states $S \in \{increase, decrease\}$ for credibility score. Then, the manufactured $\Psi_{initial}$ is sent to LLM for reasoning requests. Finally, the LLM outputs: a pair of reasoning and evaluated credibility score - $(R, V) \in \{(R^p, V^p), (R^n, V^n)\}$, where (R^p, V^p) denote positive reasoning and corresponding credibility score, and vice versa. To make the generated reasoning available to use, we design a self-reinforced reasoning rectification (SR³) method for valid reasoning generation. Algorithm 1 describes the implementation details of SR³.

Algorithm 1 takes all the inputs and outputs of Ψ as its inputs including $\{x, y, T, V_{initial}, S\}$, and $\{R^p, V^p, R^n, V^n\}$. The algorithm first checks the label of a given news, for example, if the label is fake (the conditions for fake news are in blue, and that for real news are marked with red color in the algorithm). Then, we evaluate the initially generated positive reasoning R^p (See lines 2-11). The alteration state S is confirmed as 'decrease' since positive reasoning can decrease the credibility score of fake news (See line 3). And, we set up two restrictions to inspect the quality of the generated positive reasoning in a 'While' loop, where the restrictions represent the polarity (symbolized by M) and confidence level (symbolized by I) of news, respectively (See line 4). Specifically, the polarity threshold M (initially set to 50) is used to maintain the polarity of the label-specified news x , and the confidence level I (initially set to 0) can restrict the credibility score of news being decreased while conducting positive reasoning for fake news. It is worth noting that the credibility score of fake news (or real news) is required to be decreased (or increased) as long as the expected positive reasoning R^p is returned, and the decrementation of positive reasoning-based fake news (or incrementation of positive reasoning-based real news) is restricted by the confidence level I , vice versa for real news. Moreover, the procedure of generating fake news' negative reasoning (a.k.a., y is 'fake' and T is 'negative') is illustrated in lines 12-21. In this case, the alteration state S is transformed to 'increase' aiming to decrease the confidence level of the fake news by increasing its credibility score (V^p), yet we don't demand to alter the polarity of the negative reasoning-based news (line 14), vice

Algorithm 1: Self-Reinforced Reasoning Rectification

Input: News item - x ; Label - y ; the initial credibility score - $V_{initial}$; the type of requested reasoning T ; the alteration states S for credibility score; the pair of initial positive reasoning and credibility score - (R^p, V^p) ; the pair of initial negative reasoning and score (R^n, V^n)
Parameter: Polarity threshold of credibility score - M ; Expected incrementation for confidence level - I ; Maximum number of iterations - Max_Iter
Output: The qualified reasoning $\{R^p, R^n\}$

```

1: if ( $y == 'fake' / 'real'$ ) then
2:   if  $T == 'Positive'$  then
3:      $S \leftarrow 'decrease' / 'increase'$ ;
4:     while ( $V^p > M$ ) & ( $V^p - V_{initial} > I$ ) / ' $<$ ' do
5:       Prompting  $\Psi_{re}(x, y, T, V_{initial}, S, R^p, V^p)$ ;
6:       return  $\{NewR^p, NewV^p\}$ ;
7:        $R^p \leftarrow NewR^p$ ;
8:        $V^p \leftarrow NewV^p$ ;
9:     UNTIL  $Max\_Iter$ 
10:   end while
11: end if
12: if  $T == 'Negative'$  then
13:    $S \leftarrow 'increase' / 'decrease'$ ;
14:   while ( $V^n - V_{initial} < I$ ) / ' $>$ ' do
15:     Prompting  $\Psi_{re}(x, y, T, V_{initial}, S, R^n, V^n)$ ;
16:     return  $\{NewR^n, NewV^n\}$ ;
17:      $R^n \leftarrow NewR^n$ ;
18:      $V^n \leftarrow NewV^n$ ;
19:   UNTIL  $Max\_Iter$ 
20:   end while
21: end if
22: end if
23: return  $\{R^p, R^n\}$ 

```

versa for real news. Later, the re-generated R^n will be sent back to LLMs via prompting for re-evaluating the credibility score of news until satisfying restrictions. Finally, the positive reasoning and negative reasoning that satisfy the two restrictions will be returned by the algorithm.

Datasets

In this study, we deploy three widely adopted existing fake news datasets to conduct our experiment - Politifact (Wang 2017), and Twitter-15 & 16 (Yuan et al. 2019). More specifically, we leverage our proposed SR³ to generate high-quality positive and negative reasoning for each news item from the three datasets. It is noteworthy that the labels of data items in two Twitter datasets are in four categories (i.e., non-rumors, true-rumors, false-rumors, and unverified-rumors). To adapt the settings of our method, we group the news items labeled by non-rumors and false-rumors as true samples, and the other two types of items are treated as false samples.

Methodology

Data Representations

Given a randomly selected news x and the corresponding qualified positive & negative reasoning $\{R^p, R^n\}$, it is ca-

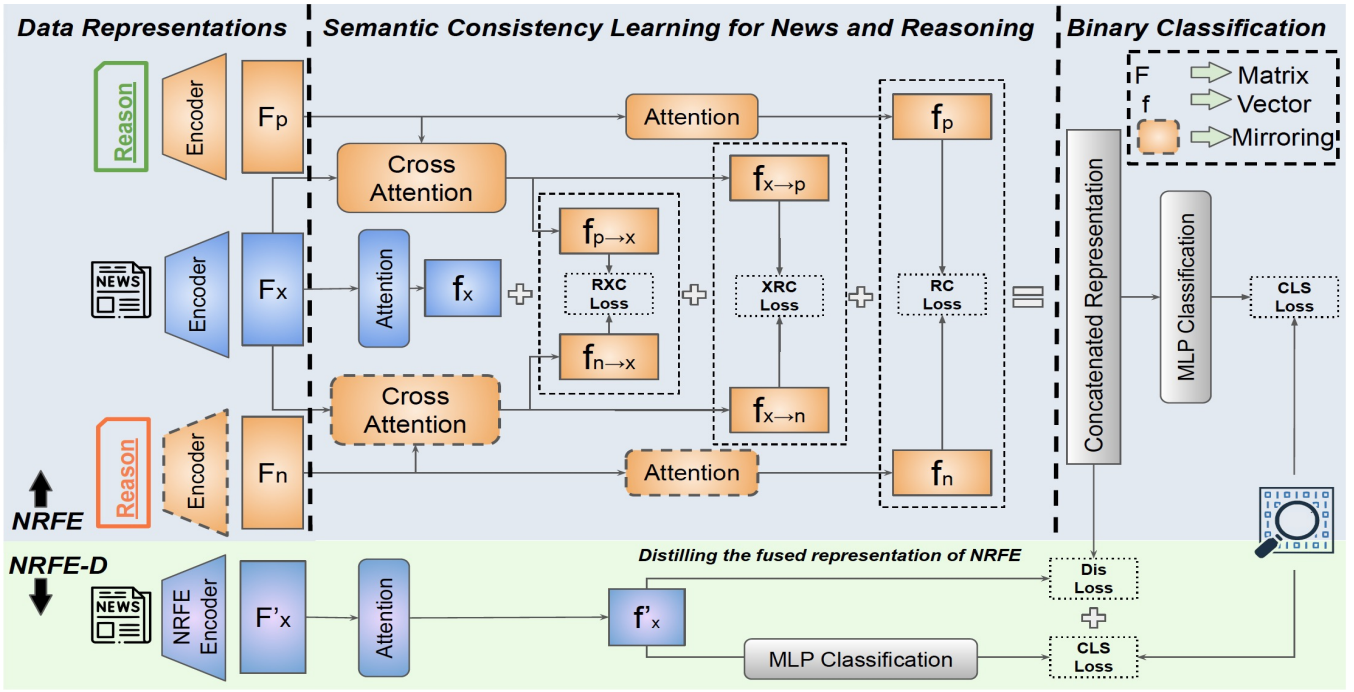


Figure 3: The system frameworks of *NRFE* and *NRFE-D*: the two Encoders embraced in *NRFE* are used to represent news and reasoning, respectively. The dashed blocks in orange color are the mirroring of the blocks in orange color, which means they are parameter shared. For the knowledge distillation model - *NRFE-D*, its Encoder used for news representation inherits the parameters of the news Encoder fine-tuned in *NRFE*. *NRFE-D* has two learning objectives: the fused representation from *NRFE* (Dis Loss: $\mathcal{L}_{dis}$) and the hard label of data (CLS Loss: $\mathcal{L}'_{cls}$).

pable of forming them as a positive news-reasoning pair - (x, R^p) and a negative pair - (x, R^n) . *NRFE* initially deploys two pre-trained BERT models as the Encoders for encoding news and reasoning, respectively. The Encoders can output sequential representations for x , and R^p (or R^n), which are denoted as F_x , and F_p (or F_n), respectively.

To investigate the relationship between news and reasoning, we leverage a cross-attention block to facilitate their collaboration. The details of the interaction between news and reasoning are as shown in Eq. 2.

$$CrossAttention \begin{cases} (F_x, F_p) = \{F_{p \rightarrow x}, F_{x \rightarrow p}\} \\ (F_x, F_n) = \{F_{n \rightarrow x}, F_{x \rightarrow n}\} \end{cases} \quad (2)$$

where $F_{p \rightarrow x}$ and $F_{n \rightarrow x}$ represent the news article x enriched with the information from positive reasoning R^p and negative reasoning R^n , respectively. In contrast, $F_{x \rightarrow p}$ (or $F_{x \rightarrow n}$) captures the relevant information from the positive reasoning (or negative reasoning) based on the queries from the news.

Semantic Consistency Learning

Since semantic gaps might exist between news and reasoning, it is necessary to transform the representations of news and reasoning into a common feature space for semantic regularization. To achieve this objective, we conduct semantic consistency learning to align the semantics

between news and reasoning. Specifically, we present three binary classification tasks to predict whether news and reasoning have a semantic correlation or not. To simplify the procedure of semantic consistency learning task between news and reasoning, we fed the seven sequential representations - F_x , F_p , F_n , $F_{p \rightarrow x}$, $F_{n \rightarrow x}$, $F_{x \rightarrow p}$, and $F_{x \rightarrow n}$ to an Attention layer to acquire their global representations - f_x , f_p , f_n , $f_{p \rightarrow x}$, $f_{n \rightarrow x}$, $f_{x \rightarrow p}$, and $f_{x \rightarrow n}$.

Given a news with corresponding positive and negative reasoning, we set the semantic correlation between news and positive reasoning as 1, and the correlation between news and negative one as 0. Then, we conduct three sub-tasks to align the semantic representations between news and reasoning - (1) the semantic consistency between news (f_x) and reasoning-aware news ($f_{p \rightarrow x}$ or $f_{n \rightarrow x}$); (2) the correlation between news (f_x) and news-aware reasoning ($f_{x \rightarrow p}$ or $f_{x \rightarrow n}$); (3) the correlation between news (f_x) and reasoning (f_p or f_n). Each of the three sub-tasks is trained on an independent MLP-based classifier. In these three sub-tasks, we adopt cosine embedding loss with a preset hyperparameter - margin d as the loss function (Eq. 3) for training.

$$\mathcal{L}_{rc} = \begin{cases} 1 - \cos(f_x, f_r), & \text{if } y = 1 \\ \max(0, \cos(f_x, f_r) - d), & \text{if } y = 0 \end{cases} \quad (3)$$

where $\mathcal{L}_{rc}$ is used to control the correlation between f_x and f_r , where $f_r \in \{f_p, f_n\}$. More specifically, $\mathcal{L}_{rc}$ can maximize the correlation between f_x and f_p , and minimize the

correlation between f_x and f_n .

Similarly, the correlation between news r_x and reasoning-aware news $f_{r \rightarrow x} \in \{f_{p \rightarrow x}, f_{n \rightarrow x}\}$ is restricted by $\mathcal{L}_{rxc}$, and the correlation between news f_x and news-aware reasoning $f_{x \rightarrow r} \in \{f_{x \rightarrow p}, f_{x \rightarrow n}\}$ is learned by $\mathcal{L}_{xrc}$. In this case, the representations of news, reasoning, and the interactions between them can be aligned into shared semantic spaces as the progress of training. Thus, the objective of semantic consistency learning $\mathcal{L}_c$ is as follows.

$$\mathcal{L}_c = \mathcal{L}_{rc} + \mathcal{L}_{rxc} + \mathcal{L}_{xrc}, \quad (4)$$

Once the training phase of the semantic correlation module is completed, the positive pair $(\{f_x, f_p\})$ is again fed into the three well-trained blocks to obtain semantically aligned representations including news representation m_x , positive reasoning representation m_p , positive reasoning-aware news representation $m_{p \rightarrow x}$, and news-aware reasoning representation $m_{x \rightarrow p}$. Finally, these four representations will be composed as an entire representation (See Eq.5), and the composition will be fed into the coming fusion module.

$$m_{final} = m_x \oplus m_p \oplus m_{p \rightarrow x} \oplus m_{x \rightarrow p} \quad (5)$$

Finally, *NRFE* take the concatenation - m_{final} as the input of an MLP to compare with the true label y , which can be formalized as follows.

$$\mathcal{L}_{cls} = CrossEntropy(y, MLP(m_{final})) \quad (6)$$

Binary Classification

Since we adopt a supervised strategy to acquire reasoning from LLMs which means the prompt used for reasoning requests maintains news labels, it is unfair to directly feed news-reasoning pairs to *NRFE* for model testing. Therefore, we design a reasoning-free model - *NRFE-D* by acquiring the information from the hidden layers of *NRFE* for knowledge distillation. Specifically, we first make *NRFE-D* employing the news Encoder fine-tuned on *NRFE* to accept news content as input for encoding the sequential representation of news F'_x . Then, the vectorized global representation of news f'_x can be obtained via an Attention layer (similar to the transformation in *NRFE*). And, f'_x goes through MLP to obtain the concatenated final representation f'_{final} . Finally, we let f'_{final} approach to the final representation m_{final} referenced from *NRFE* by leveraging reverse KL loss (Gu et al. 2024). Thus, the *NRFE-D*'s learning objective from *NRFE* is as follows (See Eq. 7).

$$\begin{aligned} \mathcal{L}_{dis} &= KL(q \parallel p) \\ &= \mathbb{E}_{x \sim p_x} \left[\sum_{m_{final}} q(f'_{final}|x) \log \frac{q(f'_{final}|x)}{p(m_{final}|x)} \right] \end{aligned} \quad (7)$$

where $p(m_{final}|x) = \frac{1}{N} \sum_{i=1}^N p(m_{final}^i|x^i)$ is the calculated probabilistic distribution from *NRFE*, and $q(f'_{final}|x)$ is the probabilistic distribution of *NRFE-D*. To reconcile the impact of misclassification in *NRFE*, we also leverage

the 'hard' label of news (i.e., y) to train *NRFE-D* using the cross-entropy loss (See Eq. 8).

$$\mathcal{L}'_{cls} = CrossEntropy(y, MLP(f'_{final})) \quad (8)$$

where the f'_{final} goes through a MLP to match the dimension of y for training *NRFE-D*. Thus, the total loss of *NRFE-D* is:

$$\mathcal{L}_D = \mathcal{L}_{dis} + \mathcal{L}'_{cls} \quad (9)$$

Experiments and Analysis

Baselines

To highlight the superiority of our proposed approach, the baselines adopted in our experiments are three-folded, (1) prompting on LLMs (LLM-only), (2) fine-tuning on pre-trained SLMs (SLM-only), and (3) task-specific fake news detection methods. Since the label of news is used to request the reasoning, all of the baselines are verified using news content only for prediction as well as *NRFE-D*.

C1: LLM-only baselines The LLMs we adopted to conduct experiments include OLLama 3 70B, Gemma 2 27B, and Mistral 7B, which are source-opened and locally deployable. The utilization of all the LLMs is under the setting of CoT and zero-shot. The performance of LLMs in fake news detection are evaluated across all the testing datasets.

C2: SLM-only baselines We deploy three Transformer Encoder-based bidirectional pre-trained language models - *BERT*, *RoBERTa*, and *ALBERT* as baselines for performance comparison. It is worth noting that these SLMs are trained on the datasets via full-parameter fine-tuning.

C3: Task-specific baselines

- **HSA** builds a hierarchical bidirectional LSTM to represent learning from the original posts and replies (Guo et al. 2018).
- **BiGRU** is one of the base models in the study (Zhang et al. 2021), which leverages BiGRU as the text Encoder to examine whether the features of publisher emotion and social emotion can facilitate the performance of fake news detection.
- **MUSER** is an evidence retrieval enhancement-based fake news detection method implemented by the sequence of summarization, retrieval, and reasoning (Liao et al. 2023).

Implementation Details

In our experiments, we set up fixed hyper-parameters including the learning rate in Adam optimizer ($3^{-5 \times 10}$), dropout rate (0.3), and the number of epochs (30 times) for both the baselines and our model. Each dataset is divided as 80% for training and 20% for testing. The metrics we adopted include accuracy (Acc.), macro F1 score (MacF1), precision (P-T for real news and P-F for fake news) as well as recall and F1 score, as shown in Tables 1-3.

Methods	PolitiFact Dataset							
	Acc.	MacF1	P-T	R-T	F1-T	P-F	R-F	F1-F
Llama 3	0.769	0.750	0.727	0.640	0.680	0.790	0.850	0.819
Gemma 2	0.622	0.509	0.518	0.186	0.274	0.639	0.892	0.744
Mistral	0.650	0.627	0.558	0.513	0.535	0.702	0.739	0.720
BERT	0.825	0.814	0.782	0.757	0.768	0.852	0.867	0.859
RoBERTa	0.810	0.794	0.784	0.702	0.736	0.829	0.877	0.851
ALBERT	0.816	0.804	0.769	0.744	0.755	0.844	0.861	0.852
HSA	0.785	0.764	0.766	0.637	0.693	0.796	0.877	0.834
BiGRU	0.740	0.714	0.702	0.592	0.631	0.770	0.833	0.797
MUSER	0.846	0.834	0.835	0.746	0.788	0.852	0.909	0.880
NRFE-D	0.857	0.847	0.831	0.792	0.808	0.876	0.897	0.885

Table 1: Performance comparison between the baselines and *NRFE-D* on the PolitiFact across multiple evaluation metrics. *NRFE-D* achieves the highest performance in almost all metrics (e.g., highest Accuracy - 0.857 and Macro F1-score - 0.847), demonstrating its robust overall performance and superiority over other baselines. MUSER and ALBERT show strong performance, but they still lag behind *NRFE-D*.

Experimental Observations

We have several observations from the results demonstrated in Tables 1-3.

First, The open-source LLMs (Llama 3, Gemma 2, Mistral) consistently generally yield the lowest results in terms of all the metrics across the three datasets compared with other methods. The phenomenon of their difficulty in adapting to fake news detection could be caused by their lack of specialization for the task-specific requirements and the potential inadequacy of their pre-training data for this specific application. More insightful, source-opened LLMs are versatile and general-purpose, but they lack task-specific training, affecting their performance in fake news detection.

Methods	Twitter-15 Dataset							
	Acc.	MacF1	P-T	R-T	F1-T	P-F	R-F	F1-F
Llama 3	0.709	0.708	0.727	0.657	0.690	0.695	0.760	0.726
Gemma 2	0.594	0.588	0.614	0.479	0.538	0.582	0.706	0.638
Mistral	0.641	0.637	0.672	0.534	0.595	0.622	0.746	0.678
BERT	0.918	0.918	0.906	0.931	0.918	0.931	0.906	0.918
RoBERTa	0.940	0.940	0.931	0.950	0.940	0.951	0.930	0.941
ALBERT	0.917	0.917	0.908	0.926	0.917	0.926	0.909	0.919
HSA	0.822	0.821	0.803	0.849	0.824	0.847	0.794	0.818
BiGRU	0.889	0.889	0.863	0.920	0.891	0.917	0.858	0.887
MUSER	0.952	0.952	0.958	0.945	0.951	0.947	0.960	0.953
NRFE-D	0.958	0.959	0.971	0.943	0.957	0.947	0.972	0.960

Table 2: Performance comparison on Twitter-15 dataset.

Second, fine-tuning on Pre-trained Language Models (BERT, RoBERTa, ALBERT) performs much better than the open-source LLMs. For example, BERT achieves high results in MacF1 (0.814) and Acc. (0.825) on PolitiFact, and

RoBERTa also demonstrates its effectiveness and robustness on two Twitter datasets. Specifically, fine-tuning these models on fake news datasets allows them to learn task-specific nuances and improve their detection capabilities.

Methods	Twitter-16 Dataset							
	Acc.	MacF1	P-T	R-T	F1-T	P-F	R-F	F1-F
Llama 3	0.771	0.769	0.853	0.729	0.786	0.690	0.828	0.753
Gemma 2	0.542	0.536	0.692	0.375	0.486	0.473	0.771	0.586
Mistral	0.650	0.649	0.743	0.604	0.666	0.568	0.714	0.632
BERT	0.920	0.914	0.955	0.895	0.924	0.868	0.942	0.904
RoBERTa	0.934	0.933	0.949	0.937	0.943	0.916	0.931	0.923
ALBERT	0.906	0.903	0.917	0.920	0.918	0.890	0.885	0.888
HSA	0.903	0.900	0.903	0.933	0.918	0.905	0.862	0.882
BiGRU	0.927	0.926	0.952	0.920	0.936	0.896	0.937	0.916
MUSER	0.938	0.939	0.957	0.937	0.947	0.916	0.942	0.929
NRFE-D	0.971	0.970	0.991	0.958	0.974	0.945	0.988	0.966

Table 3: Performance comparison on Twitter-16 dataset.

Third, the task-specific models - HSA, BiGRU, and MUSER are designed to tackle the unique challenges of fake news detection by leveraging specific aspects like social information, emotional features, and evidence retrieval. Because our datasets lack social information and emotional features, HSA and BiGRU showcase their lower performance than fine-tuning on Pre-trained Language Models, yet demonstrate their strengths compared with LLMs. MUSER demonstrates its high performance across all of the metrics (0.846, 0.952, and 0.938 in Accuracy, and 0.834, 0.952, and 0.939 in MacF1 on the three datasets), making it particularly strong in detecting misinformation.

Finally, *NRFE-D* achieves the relatively highest performance across the datasets compared with the other nine baselines (outperforms at least 1.1% Accuracy and 1.3% MacF1 on PolitiFact, 0.6% Accuracy and 0.7% MacF1 on Twitter15, and 3.3% Accuracy and 3.1% MacF1 on PolitiFact). By taking advantage of generated positive and negative reasoning via SR³, *NRFE* leverages dual BERT Encoders to learn semantic consistency between news and reasoning, thereby enriching *NRFE*'s logical and semantic comprehension for news. Based on that, *NRFE-D* not only distills the knowledge from *NRFE* but also doesn't abandon utilizing the original label of news, leading to consistent high performance across datasets and demonstrating its robustness and adaptability to diverse misinformation contexts.

Ablation Study

We conduct ablation experiments on both *NRFE* and *NRFE-D*. Specifically, we evaluate the effectiveness of the different components in our model by disabling RXC loss (i.e., w/o RXC), XRC loss (i.e., w/o XRC), RC loss (i.e., w/o RC), or RXC&XRC loss (i.e., only RC) as shown in Figs. 4 and 5.

In Fig. 4, all the different variants of *NRFE-D* yield worse performance in terms of accuracy and MacF1 compared with *NRFE-D*. We inspect the effectiveness of each block with the following observations. First, *NRFE-D* w/o RC

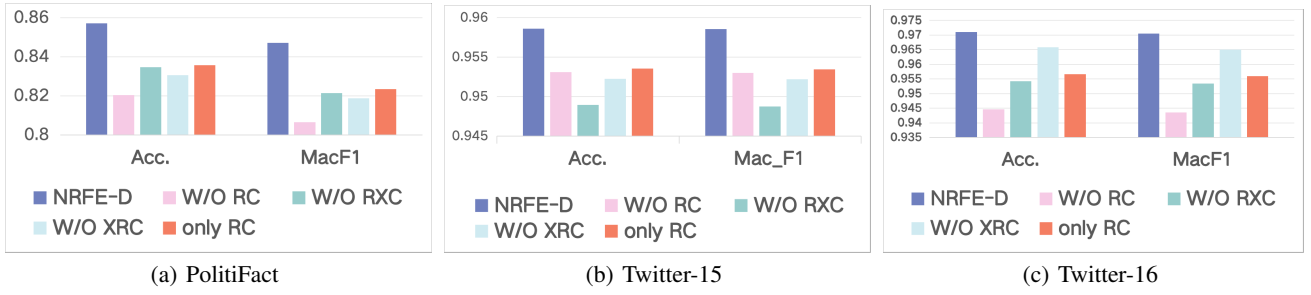


Figure 4: Ablation comparisons among *NRFE-D* and its variants in terms of accuracy and MacF1 on three datasets. RC is the RC loss used for learning the semantic correlation between news f_x and reasoning f_p or f_n as well as RXC and XRC in Fig. 3. The last variant - Only RC, denotes that RXC and XRC are disabled for ablation evaluation.

yields the relatively lowest performance on PolitiFact and Twitter-16 datasets, which demonstrates the importance of the semantic consistency between news and corresponding reasoning. In other words, *NRFE-D* ensures the semantic alignment between them by taking advantage of RC, thereby reinforcing the understanding and improving the detection of fake news. Second, w/o RXC and w/o XRC shows a bit of improvement in two metrics compared with w/o RC on PolitiFact and Twitter-16. Yet, they show a noticeable performance decline compared with *NRFE-D*. RXC and XRC can capture the interactions between news and reasoning under the cross-attention mechanism, which indicates their significance and contributions to the performance of fake news detection. Finally, only RC produces better performance than the other three variants on PolitiFact and Twitter-15, indicating the importance of RXC and XRC in *NRFE-D*.

rectness ratio in classification, which means the features of training data learned by *NRFE* are more discriminative than the others (See. Fig. 5(a)). The variant w/o RC demonstrates a significant gap between most positive and negative data samples, yet the rate of misclassification is more considerable than *NRFE* (See. Fig. 5(b)). The layout of w/o RXC is more likely a triple-classification although it has significant correctness of training classification (See. Fig. 5(c)). The variant - w/o XRC generates the worst results, which indicates the importance of news-related reasoning in affecting the performance of classification (See. Fig. 5(d)). Overall, we can see from the figures that the aggregation of features from different blocks can significantly improve the representation ability for news data, thereby enhancing the effectiveness of model training.

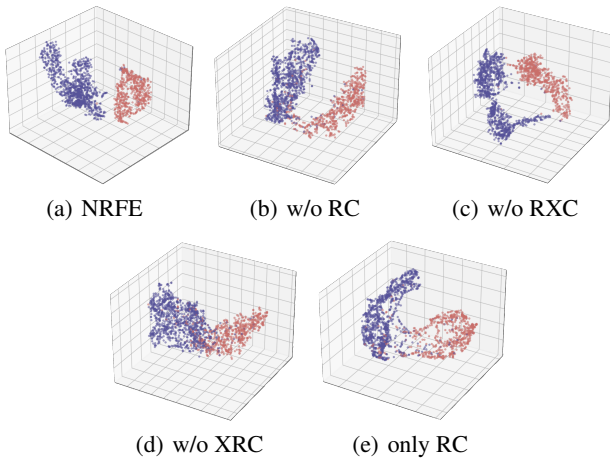


Figure 5: Ablation comparisons on PolitiFact dataset among the teacher model - *NRFE* and its four variants by visualizing their features of the last hidden layer using T-SNE. The dots in blue color denote real news, and vice versa.

Fig. 5 discover the role of different blocks in the training phase of the teacher model (i.e., *NRFE*). It is observable that the margin between fake and real news in *NRFE* is much broader than the other variants coupled with its better cor-

Conclusion

Unlike existing LLM reasoning-based fake news detection approaches, we propose a novel self-reinforced reasoning rectification method (SR^3) to conduct the production of reasoning in a supervised manner, which leverages the ability of LLMs hallucination for negative reasoning generation. In detail, we drive Ollama 3 70B to generate high-quality positive and negative reasoning for news using SR^3 via LLM reflection. Upon that, we deploy a local Negative Reasoning-based FakeE news detection model - *NRFE* to learn the semantic consistency between news and reasoning (positive or negative). To evaluate the performance of our model, we present a student model - *NRFE-D* by distilling the knowledge of *NRFE* learned from news and reasoning, thereby avoiding the impact of surprised reasoning requests. The experiment results show that our method yields outstanding performance compared with different types of SOTA baselines. In the future, we will further investigate the potentiality of other source-opened LLMs (such as Ollama 3.1, or Mistral Large 2) in generating negative reasoning in an unsupervised manner. On the other hand, LLM-based multi-agents can also be transferred in the scenes of fake news detection, which will be one of our future study directions.

Acknowledgments

We greatly appreciate the anonymous reviewers who provide insightful comments for this work. This study is partially supported by the National Natural Science Foundation of China (62403412, 62076217), the Natural Science Foundation of the Higher Education Institutions of Jiangsu Province of China under grant 23KJB520040, and the National Language Commission of China (ZDI145-71).

References

- Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; Do, Q. V.; Xu, Y.; and Fung, P. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In Park, J. C.; Arase, Y.; Hu, B.; Lu, W.; Wijaya, D.; Purwarianti, A.; and Krisnadhi, A. A., eds., *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 675–718. Nusa Dua, Bali: Association for Computational Linguistics.
- Bhattacharjee, A.; Moraffah, R.; Garland, J.; and Liu, H. 2024. Towards llm-guided causal explainability for black-box text classifiers. In *AAAI 2024 Workshop on Responsible Language Models*, Vancouver, BC, Canada.
- Boumber, D.; Tuck, B. E.; Verma, R. M.; and Qachfar, F. Z. 2024. LLMs for Explainable Few-shot Deception Detection. In *Proceedings of the 10th ACM International Workshop on Security and Privacy Analytics*, 37–47.
- Chen, C.; and Shu, K. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3): 354–368.
- Chen, Y.; Fu, Q.; Yuan, Y.; Wen, Z.; Fan, G.; Liu, D.; Zhang, D.; Li, Z.; and Xiao, Y. 2023. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 245–255.
- Gu, Y.; Dong, L.; Wei, F.; and Huang, M. 2024. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.
- Guo, H.; Cao, J.; Zhang, Y.; Guo, J.; and Li, J. 2018. Rumor detection with hierarchical social attention network. In *Proceedings of the 27th ACM international conference on information and knowledge management*, 943–951.
- Guo, H.; Zeng, W.; Tang, J.; and Zhao, X. 2023. Interpretable Fake News Detection with Graph Evidence. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 659–668.
- Hu, B.; Sheng, Q.; Cao, J.; Shi, Y.; Li, Y.; Wang, D.; and Qi, P. 2024. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. *arXiv preprint arXiv:2309.12247*.
- Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38.
- Jin, Z.; Chen, Y.; Leeb, F.; Gresele, L.; Kamal, O.; Lyu, Z.; Blin, K.; Gonzalez Adauto, F.; Kleiman-Weiner, M.; Sachan, M.; et al. 2024. Cladder: A benchmark to assess causal reasoning capabilities of language models. *Advances in Neural Information Processing Systems*, 36.
- Liao, H.; Peng, J.; Huang, Z.; Zhang, W.; Li, G.; Shu, K.; and Xie, X. 2023. MUSER: A multi-step evidence retrieval enhancement framework for fake news detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4461–4472.
- Liu, Q.; Wu, J.; Wu, S.; and Wang, L. 2024a. Out-of-distribution Evidence-aware Fake News Detection via Dual Adversarial Debiasing. *IEEE Transactions on Knowledge and Data Engineering*.
- Liu, Z.; Chen, C.; Wang, J.; Chen, M.; Wu, B.; Che, X.; Wang, D.; and Wang, Q. 2024b. Make llm a testing expert: Bringing human-like interaction to mobile gui testing via functionality-aware decisions. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 1–13.
- Ma, G.; Hu, C.; Ge, L.; Chen, J.; Zhang, H.; and Zhang, R. 2022. Towards robust false information detection on social networks with contrastive learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 1441–1450.
- Mitchell, E.; Lee, Y.; Khazatsky, A.; Manning, C. D.; and Finn, C. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, 24950–24962. PMLR.
- Pan, L.; Wu, X.; Lu, X.; Luu, A. T.; Wang, W. Y.; Kan, M.-Y.; and Nakov, P. 2023a. Fact-Checking Complex Claims with Program-Guided Reasoning. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6981–7004. Toronto, Canada: Association for Computational Linguistics.
- Pan, Y.; Pan, L.; Chen, W.; Nakov, P.; Kan, M.-Y.; and Wang, W. 2023b. On the Risk of Misinformation Pollution with Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1389–1403. Singapore: Association for Computational Linguistics.
- Pelrine, K.; Imouza, A.; Thibault, C.; Reksoprodjo, M.; Gupta, C.; Christoph, J.; Godbout, J.-F.; and Rabbany, R. 2023. Towards Reliable Misinformation Mitigation: Generalization, Uncertainty, and GPT-4. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6399–6429. Singapore: Association for Computational Linguistics.
- Qiang, J.; Li, Y.; Zhang, C.; Li, Y.; Zhu, Y.; Yuan, Y.; and Wu, X. 2023a. Chinese idiom paraphrasing. *Transactions of the Association for Computational Linguistics*, 11: 740–754.
- Qiang, J.; Liu, K.; Li, Y.; Yuan, Y.; and Zhu, Y. 2023b. ParaLS: Lexical Substitution via Pretrained Paraphraser. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for*

- Computational Linguistics (Volume 1: Long Papers)*, 3731–3746. Toronto, Canada: Association for Computational Linguistics.
- Wan, H.; Feng, S.; Tan, Z.; Wang, H.; Tsvetkov, Y.; and Luo, M. 2024. DELL: Generating Reactions and Explanations for LLM-Based Misinformation Detection. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 2637–2667. Bangkok, Thailand: Association for Computational Linguistics.
- Wang, B.; Ma, J.; Lin, H.; Yang, Z.; Yang, R.; Tian, Y.; and Chang, Y. 2024. Explainable Fake News Detection With Large Language Model via Defense Among Competing Wisdom. In *Proceedings of the ACM on Web Conference 2024*, 2452–2463.
- Wang, H.; Tang, P.; Kong, H.; Jin, Y.; Wu, C.; and Zhou, L. 2023. DHCF: Dual disentangled-view hierarchical contrastive learning for fake news detection on social media. *Information Sciences*, 645: 119323.
- Wang, W. Y. 2017. ”liar, liar pants on fire”: A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.
- Wu, Y.; Xiao, Y.; Hu, M.; Liu, M.; Wang, P.; and Liu, M. 2024. Towards Robust Evidence-Aware Fake News Detection via Improving Semantic Perception. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 16607–16618.
- Yang, J.; Jin, H.; Tang, R.; Han, X.; Feng, Q.; Jiang, H.; Zhong, S.; Yin, B.; and Hu, X. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6): 1–32.
- Ye, X.; and Durrett, G. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35: 30378–30392.
- Yin, S.; Zhu, P.; Wu, L.; Gao, C.; and Wang, Z. 2024. GAMC: an unsupervised method for fake news detection using graph autoencoder with masking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 347–355.
- Yuan, C.; Ma, Q.; Zhou, W.; Han, J.; and Hu, S. 2019. Jointly embedding the local and global relations of heterogeneous graph for rumor detection. In *2019 IEEE international conference on data mining (ICDM)*, 796–805. IEEE.
- Zeng, F.; and Gao, W. 2024. JustiLM: Few-shot Justification Generation for Explainable Fact-Checking of Real-world Claims. *Transactions of the Association for Computational Linguistics*, 12: 334–354.
- Zhang, K.; Zeng, J.; Meng, F.; Wang, Y.; Sun, S.; Bai, L.; Shen, H.; and Zhou, J. 2024. Tree-of-Reasoning Question Decomposition for Complex Question Answering with Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19560–19568.
- Zhang, X.; Cao, J.; Li, X.; Sheng, Q.; Zhong, L.; and Shu, K. 2021. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*, 3465–3476.
- Zhang, X.; and Gao, W. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *arXiv preprint arXiv:2310.00305*.