



# Explainable Fake News Detection With Large Language Model via Defense Among Competing Wisdom

Bo Wang  
School of Artificial Intelligence,  
Jilin University  
Changchun, China  
Hong Kong Baptist University  
Hong Kong SAR, China  
wangbo21@mails.jlu.edu.cn

Jing Ma\*  
Hongzhan Lin  
Hong Kong Baptist University  
Hong Kong SAR, China  
majing@comp.hkbu.edu.hk  
cshzlin@comp.hkbu.edu.hk

Zhiwei Yang  
Guangdong Institute of Smart  
Education,  
Jinan University  
Guangzhou, China  
yangzw18@mails.jlu.edu.cn

Ruichao Yang  
Hong Kong Baptist University  
Hong Kong SAR, China  
csrcyang@comp.hkbu.edu.hk

Yuan Tian\*  
School of Artificial Intelligence,  
Jilin University  
Changchun, China  
yuantian@jlu.edu.cn

Yi Chang\*  
School of Artificial Intelligence,  
Jilin University  
Changchun, China  
yichang@jlu.edu.cn

## ABSTRACT

Most fake news detection methods learn latent feature representations based on neural networks, which makes them black boxes to classify a piece of news without giving any justification. Existing explainable systems generate veracity justifications from investigative journalism, which suffer from debunking delayed and low efficiency. Recent studies simply assume that the justification is equivalent to the majority opinions expressed in the wisdom of crowds. However, the opinions typically contain some inaccurate or biased information since the wisdom of crowds is uncensored. To detect fake news from a sea of diverse, crowded and even competing narratives, in this paper, we propose a novel defense-based explainable fake news detection framework. Specifically, we first propose an evidence extraction module to split the wisdom of crowds into two competing parties and respectively detect salient evidences. To gain concise insights from evidences, we then design a prompt-based module that utilizes a large language model to generate justifications by inferring reasons towards two possible veracities. Finally, we propose a defense-based inference module to determine veracity via modeling the defense among these justifications. Extensive experiments conducted on two real-world benchmarks demonstrate that our proposed method outperforms state-of-the-art baselines in terms of fake news detection and provides high-quality justifications.

## CCS CONCEPTS

• **Information systems** → **Multimedia information systems**.

\*Joint Corresponding Author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

WWW '24, May 13–17, 2024, Singapore, Singapore

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0171-9/24/05...\$15.00

<https://doi.org/10.1145/3589334.3645471>

## KEYWORDS

Fake News Detection, Explainable, Large Language Model, Competition in Wisdom, Defense-based Inference

### ACM Reference Format:

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, Yi Chang. 2024. Explainable Fake News Detection With Large Language Model via Defense Among Competing Wisdom. In *Proceedings of the ACM Web Conference 2024 (WWW '24)*, May 13–17, 2024, Singapore, Singapore. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3589334.3645471>

## 1 INTRODUCTION

The proliferation of fake news on social media has become a remarkable concern, leading to detrimental effects on individuals and society. For example, during the global COVID-19 pandemic, a piece of spurious news claiming that “*the COVID-19 vaccine can induce serious side effects and potentially result in death*”<sup>1</sup> attracted the public’s attention, leading to people’s hesitancy and resistance towards vaccine uptake, and thus seriously impacted the virus containment effort and overwhelmed healthcare systems around the world. Fortunately, as the truth was consistently justified by the official media and investigative journalism, the public recognized the claim as fake. This indicates the positive role of solid justification in restricting the social harmfulness caused by fake news. However, relying solely on investigative journalism to enable the public to detect fake news is not a realistic approach. Such a labor-intensive and time-consuming manner limits the coverage and is debunking delayed. Thus, it is urgent to develop automated methods to detect fake news and provide clear justifications timely.

Most previous works for detecting fake news focus on incorporating various information to learn the latent features of fake news via deep networks, e.g., credibility [27], stances [22], propagation patterns [30], extra knowledge [10], out-of-domain data [18], etc. Despite their success in detecting, limited by their black-box nature, they are unable to provide any justification, which is essential to the public. To address this problem, some studies are dedicated to explainable fake news detection (EFND) that aims to generate

<sup>1</sup><https://www.bbc.com/news/53525002>

<b>Claim:</b> After the discharge of nuclear-contaminated water, there won't be any healthy salt left for humans to consume. <b>False</b>
<b>R1:</b> Nuclear-contaminated water will pollute seawater and cause salt to cause cancer. It's better to stock up on some healthy and safe salt while we can. ... <b>[support]</b>
<b>R2:</b> Damn it! I'm going to buy salt! ... <b>[support]</b>
<b>R3:</b> Don't spread panic! In the current global salt production, rock salt accounts for 41%, underground brine and salt lakes account for 29%, and sea salt accounts for 26%. Even if nuclear-contaminated water has an impact on sea salt, humans still have other sources of salt to consume. ... <b>[refute]</b>
<b>R4:</b> The presence of nuclear contaminated water increases the risk of salt-induced cancer. ... <b>[support]</b>

**Figure 1: A false claim from the Sina Weibo. The comparison of informativeness and soundness between two competing parties serves as an indicator of veracity.**

veracity prediction and explanations at the same time [15]. Many works provide their interpretability by highlighting salient words, phrases, or sentences in relevant reports via attention mechanism [23, 28, 31, 37]. However, these works only uncover regions with high contributions for the final prediction, lacking intuitive and credible explanations. As human justification brings great improvement in veracity prediction [1], some works use debunked reports to generate explanations. Atanasova et al. [3] obtains explanations by summarizing from debunked reports, but suffers from debunking delayed and low efficiency. To mitigate this problem, motivated by the effectiveness of the wisdom of crowds in fact-checking [2], Yang et al. [40] assumes the majority of opinions expressed in the raw reports are equivalent to the justifications and extracts explanations from them. However, unverified raw reports typically contain inaccurate or biased information since the wisdom of crowds is uncensored. The ill-considered assumption leads to misleading results that are biased towards the majority opinion in raw reports. Therefore, it remains a challenge to effectively leverage the rich wisdom expressed in raw reports to support EFND.

Recent detection in the field of stance detection implicitly suggests that the different insights in various raw reports are crucial signals in the quest for truth [14, 22, 39]. Inspired by it, we propose to split the wisdom into two distinct parties, which allows the detection to rely on the quality of wisdom rather than its quantity. Take a concrete example, as shown in Figure 1, there are two competing parties to the claim. For the supporting party, R1 and R4 both briefly discuss the risk of salt leading to cancer, and R2 echoes the claim without additional information. In contrast, R3 provides the refuting party with detailed evidence to illustrate its unique viewpoint, which is solid and persuasive. Based on the observation, we assume that the reports indicating truthfulness could exhibit higher quality of informativeness and soundness compared to those conveying inaccurate information. As a result, the veracity of news can be ascertained through a comparative analysis between two competing parties. Therefore, how to effectively split the wisdom into two parties from raw reports and then capture their quality divergence, is a critical problem for enhancing EFND.

To deal with the above issues, we propose a defense-based explainable fake news detection framework, which strives to capture the divergence between the competing wisdom reflected in raw

reports and pursue the veracity of claims in a defense-like way. Specifically, we first propose an evidence extraction module to split the wisdom of crowds into two competing parties, from which we detect salient evidences, respectively. Since the wisdom of each competing party is massive, it is formidable to identify the divergence by directly using the competing evidences, thereby raising the demand for streamlined summarization. Inspired by the dominating performance of large language models (LLMs) [24, 25, 33, 34], based on the respective evidences, we then design a prompt-based module to generate justifications by inferring reasons towards two possible veracities. Benefiting from the powerful reasoning and generating capacity of LLMs, we obtain the summarized wisdom of both parties in natural language, allowing for explicit comparisons. Finally, to capture the winner in quality comparisons, namely the party indicating the veracity of the claim, we propose a defense-based inference module to determine veracity by modeling the defense among these justifications. In this manner, the final justification for the verdict is adaptively selected from these justifications. Our main contributions are summarized as follows:

- We develop a defense-based framework to utilize the rich competing wisdom naturally contained in raw reports, mitigating the majority bias problem from which existing works suffer.
- By integrating the powerful reasoning capabilities of LLMs, our model can derive explanations comparable to those of human experts without any debunked reports' supervision.
- We achieve new state-of-the-art results on two real-world explainable fake news detection datasets and demonstrate the quality of the explanations with extensive analyses.

## 2 RELATED WORK

### 2.1 Explainable Fake News Detection

Much effort has been devoted to investigating the field of EFND in previous studies. To bring about some explainability, some works explore attention mechanisms to highlight salient phrases [28, 37], news attributes [38], and suspicious users [20]. In order to gain more human-readable explanations, some works capture the salient sentences as explanations via sentence-level attention [21, 23, 31]. However, these methods merely uncover regions with high contributions to the final veracity prediction rather than view the explanation generation as a dependent task. To address it, Kotonya and Toni [16] regards the explanation generation task as a pre-trained extractive-abstractive summarization task. Atanasova et al. [3] formats the EFND task as a multi-task learning problem, and then trains a joint model based on external debunked reports collected from fact-checking websites. However, the debunking of claims is a labor-intensive and time-consuming process. Such a heavy reliance on debunked reports results in coverage limitations and debunked delays. To alleviate this problem, Yang et al. [40] merely employs the debunked report as a supervised signal in training, and concentrates on the majority opinions from crowds expressed in relevant raw reports to aid in prediction and extract evidence. However, it ignores the inaccurate and biased information in unverified reports, causing a misleading result. Therefore, by splitting the wisdom of crowds into two competing parties and introducing a defense-like strategy, we capture and leverage the divergence between the competing wisdom to reach a verdict on the claim.

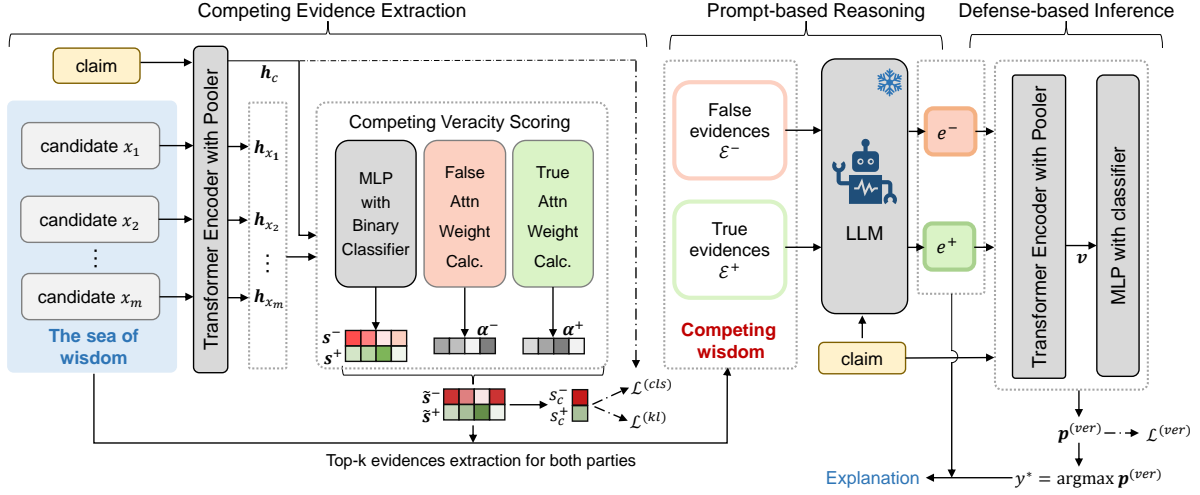


Figure 2: An overview of the proposed LLM-equipped defense-based explainable fake news detection (L-Defense) framework.

## 2.2 Large Language Model in Fake News Detection

Recently, LLMs have been proven excellent ability in multiple classification and reasoning tasks [5, 24–26, 34]. Unfortunately, the huge training cost prevents LLMs from keeping up with the latest information, which restricts the application of LLMs in the field of fake news detection with high real-time requirements. Several studies experimentally demonstrate there is a gap between LLMs and fine-tuned small language models (SLMs) like BERT [9], but also indicate that LLMs hold great potential for detecting fake news [11, 17, 32]. Li et al. [17] propose a step-by-step framework consisting of a set of plug-and-play modules to facilitate fact news detection. It achieves promising results in a zero-shot setting by purely prompting LLMs with external retrieved knowledge. Hu et al. [11] find that LLMs are suboptimal at veracity judgment but good at analyzing contents. It thus trains the small language model to adaptively acquire insights from LLM-generated rationales in a distillation framework. Cheung and Lam [7] use LoRA tuning [12] to train a LLaMA-based [33] detector with external retrieved knowledge. However, these methods mainly concentrate on the detection of fake news and lack the ability to provide an easily comprehensible explanation to the publics. Also, the methods incorporating external retrieved knowledge still suffer from the majority bias problem. In contrast, we employ LLMs to generate justifications towards two possible veracities based on respective evidences. The derived competing justifications are used to detect veracity via a defense-like process.

## 3 PROPOSED APPROACH

This section begins with a task definition of explainable fake news detection (EFND). Then, we present our LLM-equipped **Defense-based Explainable Fake News Detection (L-Defense)** framework (as in Figure 2) consisting of three components: a competing evidence extractor (§3.1), a prompt-based reasoning module (§3.2), and a defense-based inference module (§3.3). Finally, we discuss the generalization of the proposed L-Defense (§3.4).

**Task Definition.** Given a news claim  $c$  associated with a veracity label  $y$ , and its relevant raw reports  $\mathcal{D} = \{d_i\}_{i=1}^{|\mathcal{D}|}$ , where each  $d_i = (s_{i,1}, s_{i,2}, \dots, s_{i,|d_i|})$  denotes a relevant report consisting of a sequence of sentences, EFND aims to predict a veracity label  $y^*$  of claim  $c$  and provide an explanation  $e$  regarding the prediction. For different datasets,  $y^*$  can be either in the set  $\{\text{false}, \text{half}, \text{true}\}$  or  $\{\text{pants-fire}, \text{false}, \text{barely-true}, \text{half}, \text{mostly-true}, \text{true}\}$ .

### 3.1 Competing Evidence Extraction

In this subsection, we aim to split the sea of wisdom into two competing parties, and extract respective evidences for subsequent comparison. As the only available supervised data is the veracity label of the claim, we conduct a temporary veracity prediction to detect salient evidences for both parties via a veracity scoring module. Each claim is assigned a temporary veracity label in  $\{\text{false}, \text{half}, \text{true}\}$  based on its real veracity, where “half” indicates that the claim contains both true and false aspects.

**3.1.1 Claim and Evidence Representation.** Since a report may contain evidences indicating different veracities, we disaggregate the reports into individual sentences, resulting in a corpus of candidate evidence sentences  $\mathcal{X} = \{x_j\}_{j=1}^m$ , where  $m = \sum_{d_i \in \mathcal{D}} |d_i|$ . For convenience and simplicity, in the following we hide the subscripts  $j$ . We thereby adopt a vanilla pre-trained transformer encoder [35] to generate representations of claim and candidate evidences. Formally, a Transformer-Encoder is fed with a claim or a candidate,

$$h_c = \text{Pool}(\text{Transformer-Enc}(c; \theta^{(ec)})), \quad (1)$$

$$h_x = \text{Pool}(\text{Transformer-Enc}(x; \theta^{(ec)})), \quad (2)$$

where  $\text{Pool}(\cdot)$ , defined in [9], collects the resulting of [CLS] to denote a contextualized representation for the sequence,  $\theta^{(ec)}$  denotes the learnable parameters of transformer encoder.

**3.1.2 Competing Veracity Scoring.** In order to detect salient evidences for both parties, we propose to assign two veracity scores for each candidate, namely the “false score” and the “true score”,

which represent the level of support for the claim being false or true, respectively. Naturally, the two scores can be used for ranking to extract the top- $k$  “false evidences” and “true evidences” in inference.

To gain the veracity scores, we first calculate a pair of complementary scores for *true* and *false* for each candidate. Borrowing common practices from the natural language inference (NLI) area, we apply an interactive concatenation [4, 29] to the pair of claim and candidate, and then perform a neural binary classifier. Formally, we adopt the interactive concatenation written as

$$\mathbf{u} = [\mathbf{h}_c; \mathbf{h}_c \times \mathbf{h}_x; \mathbf{h}_c - \mathbf{h}_x; \mathbf{h}_x], \quad (3)$$

where  $\mathbf{u}$  is used to represent the semantic relationship between the claim and candidate. Then, a two-way classifier is applied to  $\mathbf{u}$  and produces a two-dimensional categorical distribution corresponding to the *false* and *true* veracity probabilities respectively, i.e.,

$$\mathbf{p} = P(z^{(vp)} | \mathbf{u}; \theta^{(vp)}) \triangleq \text{softmax}(\text{MLP}(\mathbf{u}; \theta^{(vp)})), \quad (4)$$

where  $\text{MLP}(\cdot)$  stands for a multi-layer perceptron, and  $\theta^{(vp)}$  is its learnable parameters. In this way, the two dimensions of  $\mathbf{p}$  can serve as *false* score  $s^-$  and *true* score  $s^+$  of the candidate to perform false and true candidate ranking, respectively.

Nevertheless, merely employing the complementary scores of each candidate for the ranking purpose is suboptimal, as not every sentence carries valuable information or contributes significantly to the final veracity prediction [3, 40]. To address this limitation, we calculate each candidate’s attention weights for the claim under two competing veracities, enabling a more precise ranking of the evidences. Formally, we present a concatenation-based attention weight calculation module to gain the false (true) attention weight,

$$\alpha^v = \frac{\exp(\text{MLP}([\mathbf{h}_c; \mathbf{h}_x]; \theta^{(v)}))}{\sum_k \exp(\text{MLP}([\mathbf{h}_c; \mathbf{h}_{x_k}]; \theta^{(v)}))}, \quad (5)$$

where  $v$  alternates between  $-$  and  $+$ ,  $\alpha^-$  and  $\alpha^+$  denote false and true weight respectively,  $\theta^{(-)}$  and  $\theta^{(+)}$  are the learnable parameters. As a result, the final *false* (*true*) score of candidate evidence used for ranking is

$$\tilde{s}^v = \alpha^v s^v. \quad (6)$$

Based on the competing veracity scores of candidate evidences, we can naturally obtain the competing veracity scores of a claim by

$$s_c^v = \sum \tilde{s}^v, \quad (7)$$

which can be used to judge the veracity of a claim.

**3.1.3 Training and Inference.** Since the only available supervised data is the veracity label of the claim, we employ two loss objectives to conduct the temporary veracity prediction, taking into consideration the extraction of competing evidences.

**Training.** The primary objective of this module is to rank and extract two competing sets of evidence using the competing veracity scores. To achieve this, we use a soft objective that considers the competing scores, utilizing the Kullback–Leibler (KL) divergence,

$$\mathcal{L}^{(kl)} = \text{KL}(\mathbf{p} || \mathbf{p}_c), \quad (8)$$

where  $\mathbf{p}_c = (s_c^-, s_c^+)$  is derived from Eq.(7),  $\mathbf{p}$  denotes the pre-defined distribution based on the temporary veracity label  $y^t$ , when

$y^t$  is *false*,  $\mathbf{p} = (1, 1)$ , when  $y^t$  is *half*,  $\mathbf{p} = (0.5, 0.5)$ , and when  $y^t$  is *true*,  $\mathbf{p} = (0, 1)$ . However, this soft KL loss is not suitable for the inference purpose and its connection with the claim is not strong enough. Hence, we further define a neural classifier for the temporary three-categorical veracity prediction as

$$\begin{aligned} \mathbf{p}^{(cls)} &= P(z^{(cls)} | \mathbf{h}_c; \mathbf{p}_c; \theta^{(cls)}) \\ &\triangleq \text{softmax}(\text{MLP}([\mathbf{h}_c; \mathbf{p}_c]; \theta^{(cls)})). \end{aligned} \quad (9)$$

Next, the classification objective to train this module is

$$\mathcal{L}^{(cls)} = - \sum_{\mathcal{D}} \log \mathbf{p}_{[\hat{y}^t=y^t]}^{(cls)}, \quad (10)$$

where  $\mathbf{p}_{[\hat{y}^t=y^t]}^{(cls)}$  denotes fetching the probability value corresponding to the temporary veracity label  $y^t$ .

We train the learnable parameters in our proposed extraction module towards a linear combination of the two losses, i.e.,

$$\mathcal{L}^{(ext)} = \gamma \mathcal{L}^{(cls)} + (1 - \gamma) \mathcal{L}^{(kl)}, \quad (11)$$

where  $\gamma$  is the trade-off weight.

**Inference.** After optimizing the extraction module w.r.t  $\mathcal{L}^{(ext)}$ ,  $\tilde{s}^-$  and  $\tilde{s}^+$  from Eq.(6) produced in inference can be used as ranking basis. For all candidate evidences, they will be ranked twice based on  $\tilde{s}^-$  and  $\tilde{s}^+$  respectively. The sets of top- $k$  false and true evidences, i.e.,  $\mathcal{E}^-$  and  $\mathcal{E}^+$ , are then prepared for the prompt-based reasoning module described in the subsequent section.

## 3.2 Prompt-based Reasoning with LLM

To effectively leverage the divergence contained in massive competing evidences on informativeness and soundness, we develop a prompt-based module for a further filter and summarization. Motivated by the remarkable abilities of LLMs in reasoning [24, 34], we engage an LLM to conduct abductive reasoning to explain why the claim is false or true based on the extracted evidence sets and a given prior veracity label.

Given a claim  $c$ , a prior label  $\tilde{y}^v$ , and an evidences set  $\mathcal{E}^v$ , to prompt the large language model in uniform language modality, we curate a template  $T$  that consists of a triplet  $\{c, \tilde{y}^v, \mathcal{E}^v\}$ . We prompt the LLM with it to generate an explanation  $e^v$  that elicits the reasoning knowledge about how to infer the veracity label  $\tilde{y}^v$  based on the interplay of the claim  $c$  and the veracity-oriented evidence  $\mathcal{E}^v$ . Specifically, we design  $T$  as:

“Given a claim:  $[c]$ , a veracity label  $[\tilde{y}^v]$ , please give me a stream-lined rationale associated with the claim, for how it is reasoned as  $[\tilde{y}^v]$ . Below are some sentences that may be helpful for the reasoning, but they are mixed with noise:  $[\mathcal{E}^v]$ . ”

The reasoning is performed for both *false* and *true* and thus two reasoning texts are obtained. As detailed previously, the evidence set which is consistent with the actual veracity of the claim brings more information and is more reasonable than the competing one. Thus, the LLM prefers to generate solid reasoning in favor of it, while providing weak reasoning with inaccurate information for its competitor, as demonstrated in Appendix D.1. In this manner, the two LLM-generated veracity-oriented reasoning  $e^-$  and  $e^+$ , which can be viewed as two explanations to clarify its relevant veracity label, will always possess a relative strength in confidence, greatly facilitating the detection of fake news.

As detailed in §2.2, LLMs can provide desirable multi-perspective rationales but still underperform the basic fine-tuned small language models [11]. Therefore, based on the generated explanations, we further propose a defense-based inference module with an SLM.

### 3.3 Defense-based Inference

With the news claim  $c$  and two veracity-oriented explanations derived from LLM, we develop a defense-based fake news detector. This detector aims to discern the relative strength of the two explanations from their defense, ultimately providing the veracity verdict. Concretely, we concatenate the three texts and feed them into a pre-trained Transformer encoder for a contextual representation. Formally,

$$\mathbf{v} = \text{Transformer-Enc}([c; [\text{SEP}]; e^-; [\text{SEP}]; e^+]; \theta^{(ed)}), \quad (12)$$

where  $[\text{SEP}]$  denotes the special separate token defined in [9]. Due to the stacked transformer encoders, this representation can effectively capture the semantic differences and connections between these three texts. Then, we define a classifier on the top of their rich representation for veracity prediction,

$$\mathbf{p}^{(ver)} = P(z^{(ver)} | \mathbf{v}; \theta^{(ver)}) \triangleq \text{softmax}(\text{MLP}(\mathbf{v}; \theta^{(ver)})). \quad (13)$$

The training objective of the detection task is written as:

$$\mathcal{L}^{(ver)} = - \sum_{\mathcal{D}} \log \mathbf{p}_{\hat{y}=y}^{(ver)}, \quad (14)$$

where  $\mathbf{p}_{\hat{y}=y}^{(ver)}$  denotes fetching the probability value corresponding to the veracity label  $y$ .

The inference procedure can be simply written as

$$y^* = \arg \max \mathbf{p}^{(ver)}. \quad (15)$$

Based on the model's predictions, we select the corresponding explanation as the final explanation  $e$ , which can be written as:

$$e = \begin{cases} e^-, & y^* = \text{false} \\ e^+, & y^* = \text{true} \\ e^-; e^+, & y^* = \text{half} \end{cases} \quad (16)$$

Especially, in the case of a "half" prediction, to help users understand the false aspect and the true aspect of the claim, we utilize a template to concatenate the two explanations and present both explanations simultaneously. This approach ensures that the final explanation aligns with the predicted veracity label.

### 3.4 Discussion about Generalization

Although our proposed model extracts competing evidences for claims used in reasoning and inference, since its specially designed three-stage architecture, it can be well-generalized to any case, regardless of the presence or absence of competing evidences. Please refer to Appendix B for more detailed discussions.

## 4 EXPERIMENT

In this section, we evaluate L-Defense on two real-world explainable fake news detection benchmarks, and verify the model's effectiveness (§4.1) and explainability (§4.2). Then, we conduct an extensive ablation study in §4.3 to verify the significance of each proposed module. Lastly, in §4.4, we do case study to show how our proposed model brings improvement.

**Table 1: Summary statistics of datasets. The numbers range from 0 to 5, representing the increasing veracity labels {pants-fire, false, barely-true, half-true, mostly-true, true}. "ALL" means the total number, and  $|S|_{avg}$  denotes the average number of sentences associated with each claim.**

		0	1	2	3	4	5	ALL	$ S _{avg}$
LIAR-RAW	train	812	1,985	1,611	2,087	1,950	1,647	10,065	62
	eval	115	259	236	244	251	169	1,274	80
	test	86	249	210	263	238	205	1,251	96
RAWFC	train	-	502	-	532	-	555	1,589	154
	eval	-	66	-	67	-	67	200	156
	test	-	66	-	67	-	67	200	157

**Datasets.** We assessed the proposed approach on two explainable datasets, i.e., *RAWFC* and *LIAR-RAW* [40], whose statistics are listed in Table 1. *RAWFC* contains the claims collected from Snopes<sup>2</sup> and relevant raw reports by retrieving claim keywords. For *LIAR-RAW*, it is extended from the public dataset LIAR-PLUS [1] with relevant raw reports, containing fine-grained claims from Politifact<sup>3</sup>. Note that we do not use any debunked justifications in the datasets for both training and inference.

**Training Setups.** We initialize the encoder in the first module (§3.1) with RoBERTa<sub>base</sub> [19] for the temporary veracity prediction, and set  $k = 10$  to extract evidences. As for the LLM used in §3.2, we alternate between ChatGPT [25] and LLaMA2<sub>7b</sub> [34]. The former refers to a widely used LLM developed by OpenAI, specifically utilizing the "gpt-3.5-turbo-0613" version, and the latter is a smaller yet powerful LLM created by Meta AI. To obtain the final prediction, we initialize the transformer encoder in §3.3 with the RoBERTa<sub>large</sub>. Please refer to Appendix C for more training details<sup>4</sup>.

### 4.1 Evaluations on Veracity Prediction

**Baselines.** We compare our L-Defense with two categories, traditional non-LLM-based approaches and LLM-based approaches. *Traditional category* contains: 1) **DEFEND** [31]; 2) **SBERT-FC** [16]; 3) **GenFE** and **GenFE-MT** [3]; 4) **CofCED** [40]. And *LLM-based category* contains: 5) **LLaMA2<sub>claim</sub>** (7b version) [34] prompts with the news claim to directly generate a veracity prediction and corresponding explanation; 6) **ChatGPT<sub>claim</sub>** [25], which is similar to LLaMA2<sub>claim</sub>; 7) **ChatGPT<sub>full</sub>** [25] prompts with the claim and all related reports, and the absence of LLaMA2<sub>full</sub> is that the 7b model struggles to produce consistent output after processing such lengthy inputs; 8) **FactLLaMA** [7] leverages the LORA tuning [12] to supervised fine-tunes a LLaMA<sub>7b</sub> with the claims; 9) **FactLLaMA<sub>know</sub>**, compared with FactLLaMA, fed with external relevant evidence retrieved from search engines.

The veracity prediction results of competitive approaches and ours on the two benchmarks are shown in Table 2. Following prior works [40], we adopt macro-averaged precision (P), recall (R), and F1 score (macF1) to evaluate the performance. It is observed that our proposed L-Defense is able to achieve state-of-the-art or competitive performance on the two datasets.

<sup>2</sup><https://www.snopes.com/>

<sup>3</sup><https://www.politifact.com/>

<sup>4</sup>The source code is available at [https://github.com/wangbo9719/L-Defense\\_EFND](https://github.com/wangbo9719/L-Defense_EFND)



**Table 2: Veracity prediction results on RAWFC and LIAR-RAW.** †Resulting numbers are reported by Yang et al. [40], and the results of FactLLaMA are taken from the original paper. The bold numbers denote the best results in each fine-grained genre while the underlined ones are state-of-the performance.

	RAWFC			LIAR-RAW		
	P	R	macF1	P	R	macF1
<i>Traditional approach</i>						
dFEND [31]†	44.93	43.26	44.07	23.09	18.56	17.51
SBERT-FC [16]†	51.06	45.92	45.51	24.09	22.07	22.19
GenFE [3]†	44.29	44.74	44.43	28.01	26.16	26.49
GenFE-MT [3]†	45.64	45.27	45.08	18.55	19.90	15.15
CofCED [40]†	<b>52.99</b>	<b>50.99</b>	<b>51.07</b>	<b>29.48</b>	<b>29.55</b>	<b>28.93</b>
<i>LLM-based approach</i>						
LLaMA2 <sub>claim</sub>	37.30	38.03	36.77	17.11	17.37	15.14
ChatGPT <sub>full</sub>	39.48	45.07	39.31	<b>29.64</b>	23.57	21.90
ChatGPT <sub>claim</sub>	<b>47.72</b>	<b>48.62</b>	<b>44.43</b>	25.41	<b>27.33</b>	<b>25.11</b>
FactLLaMA [7]	53.76	54.00	53.76	32.32	31.57	29.98
FactLLaMA <sub>know</sub> [7]	<b>56.11</b>	<b>55.50</b>	<b>55.65</b>	<b>32.46</b>	<b>32.05</b>	<b>30.44</b>
<i>Ours</i>						
L-Defense <sub>LLaMA2</sub>	60.95	60.00	60.12	<b>31.63</b>	31.71	<b>31.40</b>
L-Defense <sub>ChatGPT</sub>	<b>61.72</b>	<b>61.01</b>	<b>61.20</b>	30.55	<b>32.20</b>	30.53

For the traditional approaches, most of them underperform ChatGPT<sub>claim</sub>, demonstrating the potential of LLMs in fake news detection. For the first three LLM-based approaches without any tuning, ChatGPT<sub>claim</sub> achieves the best results. LLaMA2<sub>claim</sub> loses as its model size is significantly smaller than that of ChatGPT. And a possible reason why ChatGPT<sub>full</sub> loses is that the LLM is easily biased by the massive input reports. Despite the good performance of ChatGPT<sub>claim</sub>, it falls short when compared to the best method in the traditional approach, namely CofCED. By contrast, the fine-tuned LLM-based model, i.e., FactLLaMA and FactLLaMA<sub>know</sub> achieves the best results in addition to our proposed model. This indicates that simply utilizing LLMs for inference yields limited results, while carefully considering how to further leverage LLMs can lead to improved performance.

Our proposed model makes use of LLM as a reasoner in a novel defense-based framework, and then achieves excellent results on veracity prediction. The improvement is especially significant on RAWFC. Compared with the FactLLaMA<sub>know</sub>, both versions of our model achieved at least a 4% enhancement across all metrics. And on LIAR-RAW, although our model achieves inferior results on precision, both variants consistently outperform in terms of macF1. Furthermore, in comparison with ours LLaMA2 variant, ours ChatGPT variant achieves slightly superior results on RAWFC while maintaining competitiveness on LIAR-RAW, which suggests that our model does not prioritize the size of the LLM component. A possible reason of this discrepancy is that the  $|S|_{avg}$  of the two datasets differs significantly, resulting in RAWFC being more information-rich than LIAR-RAW. Consequently, with a much larger scale than LLaMA2, ChatGPT adds more additional information for reasoning, which aligns with the average lengths of explanations as listed in Table 9 in Appendix D.2.

## 4.2 Evaluations on Explanation

*Evaluation Metrics.* For the evaluation of explanations, traditional automated evaluation metrics are inadequate to assess the output results of LLMs [5]. Fortunately, Chen et al. [6] demonstrates that ChatGPT excels in assessing text quality from multiple angles, even in the absence of reference texts. Also, some works reveal that ChatGPT evaluation produces results similar to expert human evaluation [8, 13]. Therefore, we engage ChatGPT to evaluate the quality of explanations based on four metrics which have been widely employed in human evaluation [36, 40]: *misleadingness*, *informativeness*, *soundness*, and *readability*. A 5-point Likert scale was employed, where 1 represented the poorest and 5 the best in addition to misleadingness. The definitions of the metrics are:

- **Misleadingness (M)** assesses whether the model’s explanation is consistent with the real veracity label of a claim, with a rating scale ranging from 1 (not misleading) to 5 (very misleading);
- **Informativeness (I)** assesses whether the explanation provides new information, such as explaining the background and additional context, with a rating scale ranging from 1 (not informative) to 5 (very informative);
- **Soundness (S)** describes whether the explanation seems valid and logical, with a rating scale ranging from 1 (not sound) to 5 (very sound);
- **Readability (R)** evaluates whether the explanation follows proper grammar and structural rules, and whether the sentences in the explanation fit together and are easy to follow, with a rating scale ranging from 1 (poor) to 5 (excellent).

In order to verify the effectiveness of LLM evaluation, we further propose an automated metric called **Discrepancy (D)**, which is an objective version of misleadingness and does not consider the quality of the explanation. It is obtained by calculating the absolute difference between the predicted and actual labels’ scores. Specifically, for RAWFC, the scores for the three labels are [0, 2.5, 5]; for LIAR-RAW, the scores for the six labels are [0, 1, 2, 3, 4, 5]. The larger the discrepancy between the predicted and true labels, the higher the score, indicating a greater degree of misleading.

*Baselines.* Observed on the veracity prediction results in Table 2, we propose the following baselines: 1) **Oracle** generates an explanation for why the claim is classified as its actual veracity label, by providing both the claim and the actual veracity label to ChatGPT; 2) **CofCED** [40], the best model in traditional approaches; 3) **ChatGPT<sub>claim</sub>**, which performs better than the LLaMA2 version; 4) **ChatGPT<sub>full</sub>**. To make a fair comparison, we limited the number of extracted sentences for CofCED. The length of explanations generated by each model can be found in the Appendix D.2. To gain the explanations of our proposed model on LIAR-RAW, we categorize *pants-fire*, *false*, and *barely-true* as *false*, *half-true* remains as *half-true*, and *mostly-true* and *true* are viewed as true. As a result, the explanation of each veracity can be derived based on Eq.(16).

The evaluation results for the quality of explanations are presented in Table 3, showing that L-Defense consistently achieves excellent performance. The alignment of score trends on  $M$  and  $D$  provides partial validation of the effectiveness of our evaluation methodology. The Oracle prompted with the actual veracity achieves superiority across almost all metrics, which can be

**Table 3: Evaluation results of explanation quality using a 5-Point Likert scale rating by ChatGPT on RAWFC and LIAR-RAW. For metrics D and M, a lower score indicates better performance, while a higher score indicates better performance for the remaining metrics. The bold numbers denote the best results in addition to Oracle while the underlined ones are better than Oracle.**

	RAWFC					LIAR-RAW				
	D	M	I	S	R	D	M	I	S	R
Oracle	-	1.52	4.46	4.73	4.72	-	1.85	4.44	4.60	4.69
CofCED [40]	1.53	2.74	2.89	1.93	2.46	1.33	3.64	1.75	1.76	1.59
ChatGPT <sub>full</sub>	1.81	2.07	<b>4.44</b>	4.62	<b>4.69</b>	1.39	2.29	3.71	4.04	3.99
ChatGPT <sub>claim</sub>	1.70	1.97	<u>4.00</u>	4.44	4.68	1.39	2.27	3.93	4.29	4.50
L-Defense <sub>LLaMA2</sub>	<b>1.30</b>	1.95	<b>4.44</b>	<b>4.67</b>	4.62	1.36	2.20	<b>4.39</b>	<b>4.64</b>	<b>4.63</b>
L-Defense <sub>ChatGPT</sub>	<b>1.30</b>	<b>1.91</b>	4.17	4.41	4.49	<b>1.31</b>	<b>2.06</b>	4.12	4.28	4.47

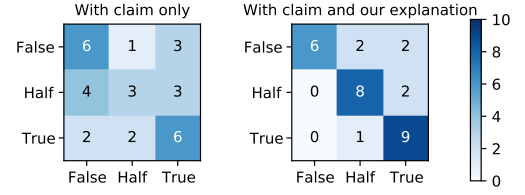
**Table 4: Explanation evaluation results using a 5-Point Likert scale rating by both ten human annotators and ChatGPT on 30 randomly sampled samples from RAWFC’s testset. Scores from ten annotators were averaged. The bold numbers denote the best results in addition to Oracle.**

	ChatGPT				Human			
	M	I	S	R	M	I	S	R
Oracle	1.53	4.50	4.77	4.77	1.47	3.61	3.89	3.86
CofCED [40]	2.90	2.77	2.87	2.47	2.46	2.91	2.47	2.44
ChatGPT <sub>full</sub>	2.07	4.43	<b>4.67</b>	<b>4.73</b>	2.22	3.22	3.38	<b>3.57</b>
ChatGPT <sub>claim</sub>	2.33	4.17	4.43	4.63	2.68	2.68	2.84	3.27
L-Defense <sub>LLaMA2</sub>	1.87	<b>4.50</b>	<b>4.67</b>	4.67	2.12	3.48	3.37	3.49
L-Defense <sub>ChatGPT</sub>	<b>1.77</b>	4.40	4.60	4.53	<b>1.97</b>	<b>3.68</b>	<b>3.52</b>	3.56

viewed as a ceiling of explanation quality. Compared to it, L-Defense achieves superior or comparable performance, highlighting the excellence of our method in generating human-read explanations.

Since the CofCED generates explanations by extracting from reports, the discrete evidence sentences are hard to fit together and may overlap. Therefore, its explanation achieves the worst performance compared with other LLM-based models. Also, its incoherent explanation leads to a high misleadingness score, though the discrepancy score is much better. This demonstrates the significance of coherence in generating a streamlined and understandable explanation. In terms of two versions of ChatGPT, ChatGPT<sub>full</sub> exhibits better performance than ChatGPT<sub>claim</sub> on RAWFC for the last three metrics. However, it underperforms ChatGPT<sub>claim</sub> on LIAR-RAW across all metrics. One possible reason is that in the full version, the input length for RAWFC is approximately twice that of LIAR-RAW, thereby incorporating more information. At the same time, this also indicates that ChatGPT is highly vulnerable to the input context. The introduction of additional raw reports, in contrast to ChatGPT<sub>claim</sub>, leads to a completely different performance in ChatGPT<sub>full</sub>. For the two variants of our proposed L-Defense, the ChatGPT variant always beats the LLaMA2 variant on misleading-related metrics while losing in the latter three metrics. It further demonstrates that the performance of our model is not limited by the size of the LLM.

*Human Evaluation.* To further validate the effectiveness of the LLMs evaluation and the helpfulness of our proposed model, we



**Figure 3: Confusion matrixes of the judgment results made by 10 annotators on 30 randomly sampled samples. The left one is derived by providing annotators with only the claim, while the right one is derived by providing annotators with both the claim and the explanations generated by our L-Defense. The results from 10 annotators were averaged and rounded off.**

**Table 5: Ablation study of veracity prediction on RAWFC.**

Method	P	R	F1
L-Defense <sub>LLaMA2</sub>	60.95	60.00	60.12
w/o evidences	54.45	52.56	52.51
with random evidences	57.09	56.47	56.35
w/o prior label	55.97	56.02	55.98
w/o explanations	52.92	51.96	51.83
w/o inference training	39.30	38.88	29.71

asked ten English-speaking adult volunteers as annotators and then conducted two human evaluations for the explanation quality. On the one hand, we instructed the annotators to perform similar evaluations as ChatGPT did. As shown in Table 4, the evaluative results of the annotators on various metrics are largely consistent with the ranks made by ChatGPT. The main difference is that humans prefer the explanations made by the ChatGPT variant of our model while ChatGPT prefers the LLaMA2 variant. On the other hand, as shown in Figure 3, our proposed model remarkably reduces error judgments and aids humans in understanding truth.

### 4.3 Ablation Study

To evaluation the contribution of each component, we conduct an extensive ablation study for L-Defense<sub>LLaMA2</sub> on veracity prediction by removing or replacing the key component: 1) “w/o evidences” removes the extracted competing evidences  $\mathcal{E}^v$  from the prompt template  $T$  for LLM; 2) “with random evidence” replaces the  $\mathcal{E}^v$  with random sampled sentences in  $T$ ; 3) “w/o prior label” removes the given prior veracity label  $\tilde{y}^v$  in  $T$ ; 4) “w/o explanations” replaces the LLM-generated veracity-oriented explanations  $e^v$  with corresponding extracted competing evidences  $\mathcal{E}^v$  in defense-based inference (§3.3); 5) “w/o inference training” replaces the fine-tune process (§3.3) with ChatGPT’s predictions.

As shown in Table 5, it is observed that the original versions significantly outperform their component-deprived versions. The degradation of “w/o evidences” and “with random evidences” demonstrate the necessity of the evidence extraction module. Moreover, despite the “with random evidences” being inferior to L-Defense, it still achieved commendable performance. This indicates the superiority of our proposed defense-based framework and the powerful ability of LLM. Furthermore, the results of “w/o prior label” argue the importance of an enforced prior label to instruct an LLM.

**Table 6: Case study. The claim is selected from RAWFC testset. Four methods’ predictions and explanations are listed. The “Gold” method denotes the judgement from Snopes. We display the top-3 evidences extracted by our extraction module and CofCED. The fragments supporting the claim as true are in green background, while the fragments supporting the claim as false are in red background.**

<b>Claim:</b> Former President Barack Obama’s administration was to blame for the shortage of protective equipment like N95 respirator masks in the early months of the 2020 COVID-19 pandemic.	
<b>Method:</b> Gold, <b>Prediction:</b> half	
· What’s true: The U.S. federal stockpile of N95 protective face masks was largely depleted during the 2009 swine flu outbreak and was not restocked.	
· What’s false: However, the Obama administration was not solely responsible for the current shortage of masks. In the intervening years, the stockpile went un replenished as the Trump administration failed to heed indications that dramatic shortages could occur.	
<b>Method:</b> L-Defense <sub>ChatGPT</sub> , <b>Prediction:</b> half	
· Extracted true-supporting evidences: (1) WASHINGTON — Senior Trump administration official on Thursday blame a shortfall of N95 mask early in the coronavirus pandemic on mismanagement after the 2009 swine flu pandemic under former President Barack Obama. (2) Health care professional and expert say we now have a shortage of mask, gown, glove, ... (3) ..., a scant 1 percent of the estimate 3.5 billion mask the nation would need in a severe pandemic.	
· Extracted false-supporting evidences: (1) Mr. Trump appear intent on focus attention on the Obama administration at a time when his own handling of the outbreak have come under intense criticism. (2) And Trump attempt to blame a president who leave office more than three year ago for his failure to act two month ago might be his most pathetic blame-shirking yet. (3) Dr. Luciana Borio, ...	
· What’s true: A senior Trump administration official attributed the shortfall of N95 masks to mismanagement following the 2009 swine flu pandemic under Obama. ... the stockpile of masks was significantly below the estimated amount needed for a severe pandemic. ...	
· What’s false: The sentences provided contain statements that criticize President Trump for attempting to shift blame onto the Obama administration, but they do not provide any concrete evidence or specific actions taken by the Obama administration that would have caused the shortage. ...	
<b>Method:</b> CofCED, <b>Prediction:</b> false	
(1) The World Health Organization have say mask, goggles and other protective equipment use by health worker be in short supply due to “rise demand, hoard and misuse.” (2) “Let’s stay calm, listen to the experts, and follow the science.” tweets The death toll in the US from the virus rose to 11 on Wednesday. (3) Washington: Former US president Barack Obama call Wednesday for people to take “common sense precaution” over the coronavirus outbreak—advising them to follow hand-washing guideline but not to wear masks.	
<b>Method:</b> ChatGPT <sub>full</sub> , <b>Prediction:</b> false	
The reports provided primarily discuss the shortage of protective equipment and the lack of preparedness in general, but they do not attribute the blame solely to the Obama administration. ... They also highlight the failure of the current administration to replenish the stockpile after the H1N1 outbreak. ...	

And the results of “w/o explanations” prove the necessity of the prompt-based reasoning module. Finally, the significant performance degradation of “w/o inference training” reveals a clear gap between LLM and fine-tuned SLM in veracity prediction. Also, it highlights that the final step of our framework is indispensable.

#### 4.4 Case Study

*Does the evidence extraction module successfully separate two competing parties from the sea of wisdom?* As shown in Table 6, the top-3 true evidences and false evidences extracted by L-Defense focus on the different viewpoints of the claim. The true evidences highlight the mask shortage caused by mismanagement following the 2009 swine flu pandemic under Obama. The false evidences indicate Trump failed to pay attention to the shortage and suggest the news may stem from Trump’s influence. These two competing evidence sets are consistent with their respective gold justifications, which verifies the effectiveness of the evidence extraction module.

*Does the two competing justifications generated by the prompt-based module help to determine veracity?* Based on the extracted competing evidences, the two justifications generated from the LLM reasoning show an obvious competition on the veracity of the claim. Since their strengths in confidence are equivalent in general, the defense-based inference module gives a correct prediction.

*Does the defense-based strategy mitigate the majority bias problem?* With the same information provided, only our proposed model makes the correct prediction. The top-ranked evidences from CofCED do not contain any evidence about the true aspect and is biased by the third evidence, causing a misclassification. The

ChatGPT<sub>full</sub> gives a justification that ignores the true points and predicts the claim as false, due to the lack of deep thinking or training. In contrast, our framework avoids these problems benefiting from our novel defense-based strategy.

## 5 CONCLUSION

In this work, we propose a novel defense-based framework by effectively leveraging the competing wisdom inherent in raw reports for explainable fake news detection. Specifically, we first propose an evidence extraction module to detect salient evidence for two competing parties. Since the extracted competing evidence is diverse and massive, we then design a prompt-based module integrating the powerful reasoning ability of LLM to generate streamlined justifications for two possible veracities. Finally, we determine the veracity of a claim by modeling the defense among these justifications and give the final explanations based on the prediction. The experiments on two real-world benchmarks can greatly support our motivations, and empirical results show state-of-the-art performance with explainability.

## ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their valuable and helpful comments. This work is supported by the National Key Research and Development Program of China (No. 2023YFF0905400), the National Natural Science Foundation of China (No. U2341229, No. 62206233), the Science and Technology Development Program of Jilin Province (No. 20210508060RQ), and Hong Kong RGC ECS (No. 22200722).



## REFERENCES

- [1] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the first workshop on fact extraction and verification (FEVER)*, 85–90.
- [2] Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science advances* 7, 36 (2021), eabf4393.
- [3] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating Fact Checking Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetraault (Eds.). Association for Computational Linguistics, 7352–7364. <https://doi.org/10.18653/v1/2020.acl-main.656>
- [4] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Mar-ton (Eds.). The Association for Computational Linguistics, 632–642. <https://doi.org/10.18653/v1/d15-1075>
- [5] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A Survey on Evaluation of Large Language Models. *CoRR abs/2307.03109* (2023). <https://doi.org/10.48550/arXiv.2307.03109>
- [6] Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the Use of Large Language Models for Reference-Free Text Quality Evaluation: A Preliminary Empirical Study. *CoRR abs/2304.00723* (2023). <https://doi.org/10.48550/arXiv.2304.00723>
- [7] Tsun-Hin Cheung and Kin-Man Lam. 2023. FactLLaMA: Optimizing Instruction-Following Language Models with External Knowledge for Automated Fact-Checking. *CoRR abs/2309.00240* (2023). <https://doi.org/10.48550/arXiv.2309.00240>
- [8] David Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 15607–15631. <https://doi.org/10.18653/v1/2023.acl-long.870>
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. <https://doi.org/10.18653/v1/n19-1423>
- [10] Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2021. Kan: Knowledge-aware attention network for fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35, 81–89.
- [11] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2023. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. *CoRR abs/2309.12247* (2023). <https://doi.org/10.48550/arXiv.2309.12247>
- [12] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. <https://openreview.net/forum?id=nZeVKeeFYf9>
- [13] Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In *Companion Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, Ying Ding, Jie Tang, Juan F. Sequeda, Lora Aroyo, Carlos Castillo, and Geert-Jan Houben (Eds.). ACM, 294–297. <https://doi.org/10.1145/3543873.3587368>
- [14] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.
- [15] Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking: A Survey. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, Donia Scott, Núria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, 5430–5443. <https://doi.org/10.18653/v1/2020.coling-main.474>
- [16] Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking for Public Health Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 7740–7754. <https://doi.org/10.18653/v1/2020.emnlp-main.623>
- [17] Miaoan Li, Baolin Peng, and Zhu Zhang. 2023. Self-Checker: Plug-and-Play Modules for Fact-Checking with Large Language Models. *CoRR abs/2305.14623* (2023). <https://doi.org/10.48550/arXiv.2305.14623>
- [18] Hongzhan Lin, Jing Ma, Liangliang Chen, Zhiwei Yang, Mingfei Cheng, and Guang Chen. 2022. Detect Rumors in Microblog Posts for Low-Resource Domains via Adversarial Contrastive Learning. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, 2543–2556. <https://doi.org/10.18653/V1/2022.FINDINGS-NAACL.194>
- [19] 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. *CoRR abs/1907.11692* (2019). <https://arxiv.org/abs/1907.11692>
- [20] Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetraault (Eds.). Association for Computational Linguistics, 505–514. <https://doi.org/10.18653/v1/2020.acl-main.48>
- [21] Jing Ma, Wei Gao, Shafiq R. Joty, and Kam-Fai Wong. 2019. Sentence-Level Evidence Embedding for Claim Verification with Hierarchical Attention Networks. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 2561–2571. <https://doi.org/10.18653/v1/p19-1244>
- [22] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Detect Rumor and Stance Jointly by Neural Multi-task Learning. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon, France, April 23-27, 2018*, Pierre-Antoine Champin, Fabien Gandon, Mounia Lahmas, and Panagiotis G. Ipeirotis (Eds.). ACM, 585–593. <https://doi.org/10.1145/3184558.3188729>
- [23] Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining Fact Extraction and Verification with Neural Semantic Matching Networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 6859–6866. <https://doi.org/10.1609/aaai.v33i01.33016859>
- [24] OpenAI. 2023. GPT-4 Technical Report. *CoRR abs/2303.08774* (2023). <https://doi.org/10.48550/arXiv.2303.08774>
- [25] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*. [http://papers.nips.cc/paper\\_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html)
- [26] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*. [http://papers.nips.cc/paper\\_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html)
- [27] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*. 1003–1012.
- [28] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 22–32. <https://doi.org/10.18653/v1/d18-1003>
- [29] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3980–3990. <https://doi.org/10.18653/v1/D19-1410>
- [30] Yuxiang Ren, Bo Wang, Jiawei Zhang, and Yi Chang. 2020. Adversarial Active Learning based Heterogeneous Graph Neural Network for Fake News Detection. In *20th IEEE International Conference on Data Mining, ICDM 2020, Sorrento,*

- Italy, November 17–20, 2020, Claudia Plant, Haixun Wang, Alfredo Cuzzocrea, Carlo Zaniolo, and Xindong Wu (Eds.). IEEE, 452–461. <https://doi.org/10.1109/ICDM50108.2020.00054>
- [31] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. dFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4–8, 2019*, Ankur Teredesai, Vipin Kumar, Ying Li, Römer Rosales, Evimaria Terzi, and George Karypis (Eds.). ACM, 395–405. <https://doi.org/10.1145/3292500.3330935>
- [32] Jinyan Su, Terry Yue Zhuo, Jonibek Mansurov, Di Wang, and Preslav Nakov. 2023. Fake News Detectors are Biased against Texts Generated by Large Language Models. *CoRR abs/2309.08674* (2023). <https://doi.org/10.48550/arXiv.2309.08674> arXiv:2309.08674
- [33] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR abs/2302.13971* (2023). <https://doi.org/10.48550/arXiv.2302.13971> arXiv:2302.13971
- [34] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrubhi Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
- [35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [36] Han Wang, Ming Shan Hee, Md. Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2023. Evaluating GPT-3 Generated Explanations for Hateful Content Moderation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th–25th August 2023, Macao, SAR, China*. ijcai.org, 6255–6263. <https://doi.org/10.24963/ijcai.2023/694>
- [37] Lianwei Wu, Yuan Rao, Ling Sun, and Wangbo He. 2021. Evidence Inference Networks for Interpretable Claim Verification. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2–9, 2021*. AAAI Press, 14058–14066. <https://doi.org/10.1609/aaai.v35i16.17655>
- [38] Fan Yang, Shiva K. Pentiyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D. Ragan, Shuiwang Ji, and Xia (Ben) Hu. 2019. XFake: Explainable Fake News Detector with Visualizations. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13–17, 2019*, Ling Liu, Ryan W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 3600–3604. <https://doi.org/10.1145/3308558.3314119>
- [39] Ruichao Yang, Jing Ma, Hongzhan Lin, and Wei Gao. 2022. A Weakly Supervised Propagation Model for Rumor Verification and Stance Detection with Multiple Instance Learning. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11–15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 1761–1772. <https://doi.org/10.1145/3477495.3531930>
- [40] Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022. A Coarse-to-fine Cascaded Evidence-Distillation Neural Network for Explainable Fake News Detection. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12–17, 2022*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, 2608–2621. <https://aclanthology.org/2022.coling-1.230>

## A THE MEASURING OF MAJORITY BIAS

To measure the majority bias, for each claim, we calculate the ratio of sentences among all candidate sentences that are semantically similar to the extracted evidences, which can be written as:

$$r = \frac{1}{k} \sum_{i=1}^k \left( \frac{1}{m} \sum_{j=1}^m \mathbb{I}(\cos(\mathbf{h}_i, \mathbf{h}_j) > 0.5) \right),$$

where  $k$  denotes the number of extracted evidences,  $m$  denotes the number of all candidate sentences,  $\mathbb{I}(\cdot)$  is an indicator function that returns 1 when the condition in  $\cdot$  is true and 0 otherwise,

$\mathbf{h}_i$  and  $\mathbf{h}_j$  denote the representations of extracted evidence and candidate sentence derived from SBERT [29] respectively,  $\cos(\cdot)$  denotes the cosine similarity computation, and 0.5 is a threshold. A larger value of this ratio indicates a higher degree of suffering from majority bias. Based on the test set of RAWFC, the average ratios of our main baseline CofCED [40] and our L-Defense are 0.32 and 0.22 respectively (we will add this result in the paper revision). Moreover, as shown in Table 4, the results of human evaluation on the Informativeness (I) also suggest that the evidences extracted by CofCED are much less informative compared with ours, since the CofCED easily falls into a narrow perspective, i.e., the major perspective. To handle this bias, our proposed defense-based approach provides an effective solution.

## B DETAILED DISCUSSION ABOUT GENERALIZATION

As we detailed in §3, the proposed L-Defense is a three-stage model. For the first stage, the competing evidence extraction module assigns two competing scores for each candidate evidence and then ranks the true-oriented and false-oriented evidence sets respectively. Thus, whether a claim contains competing evidences or not, the two evidence sets can be obtained. If a claim has no competing evidences, there are two cases: (1) no evidences supports the claim as true, the stances of all evidences are mixed with the false-oriented and neutral (i.e., no-true), and (2) no evidences supports the claim as false, the stances of all evidences are mixed with the true-oriented and neutral (i.e., no-false). For the “no-true” (or “no-false”) case, regardless of the performance of the extraction module, both evidence sets contain evidences to support its false (or true) veracity or neutral. In the second stage, since a prior veracity label is used to prompt the LLM, the LLM will give a rationale according to the label (§3.2). Without true-oriented (or false-oriented) evidences, the LLM would generate inconsistent or untenable rationales for the true-oriented (or false-oriented) party. In the third stage, based on the obvious differences between the two competing explanations in terms of soundness and informativeness, the well-trained defense-based inference module will easily make an accurate prediction (§3.3). Therefore, our model can be applied to the claims with/without competing evidences.

## C IMPLEMENTATION DETAILS

### C.1 Training Setup

For the two trained modules, we use a mini-batch Stochastic Gradient Descent (SGD) to minimize the loss functions, with Adam optimizer, 10% warm-up, and a linear decay of the learning rate. Other hyper-parameters used in training are listed in Table 7. For the training of evidence extractor on LIAR-RAW, the temporary veracity labels are assigned by categorizing *pants-fire*, *false*, and *barely-true* as false, *half-true* as half-true, and *mostly-true* and *true* as true. For the prompt-based reasoning, we set the temperature to 0.8, allowing the LLM to flexibly apply the provided evidence and its own knowledge for a rich justification. And we set the temperature as 0 during the explanation evaluation. Veracity prediction results are the best values from ten runs. All experiments were conducted using a single A40 GPU.

## C.2 Prompt Designing

For the system prompt to the ChatGPT and LLaMA2 in the proposed prompt-based module, we design the message as: For the system prompt to the ChatGPT and LLaMA2 in the proposed prompt-based module, we designed it with the fundamental concept of addressing three key aspects: *What are you? What should you do? And what is your goal?* For the user prompt, we integrate all needed elements into the instruction. Following the initial design, we used these prompts to check the format of the LLM’s output by testing them on randomly selected samples from the training set. Our criterion for success was to obtain a streamlined rationale that enables readers to assess the claim’s veracity without relying on additional background knowledge. Based on the LLM’s output, we iteratively refined the prompts through several rounds of revisions until arriving at the current version. The final user prompt we used is listed in the in §3.2. And the final system prompt is:

*“You have been specially designed to perform abductive reasoning for the fake news detection task. Your primary function is that, according to a veracity label about a news claim and some sentences related to the claim, please provide a streamlined rationale, for how it is reasoned as the given veracity label. Note that the related sentences may be helpful for the explanation, but they are mixed with noise. Thus, the rationale you provided may not necessarily need to rely entirely on the sentences above, and there is no need to explicitly mention which sentence was referenced in your explanation. Your goal is to output a streamlined rationale that allows people to determine the veracity of the claim when they read it, without requiring any additional background knowledge. The length of your explanation should be less than 200 words.”*

## D MORE EXPERIMENTS

### D.1 Validation of Assumption

Whether the assumption that the party indicating the truth is more informative and sounder than its competitors is true. As shown in Table 8, in a quantitative view, the comparable results between the true and false justifications across different classes support our assumption. When the claim is false, the evaluation results of false justification are better than those of the true one; when the claim is half-true, both yield competitive results; and when it is true, the true-oriented one is better.

### D.2 The Length of Explanations

We list the average length of explanations generated by ours and all baselines in Table 9. For a fair comparison, we selected the top-6 sentences ranked by CofCED [40] as its explanation. We limited the generated length of other LLM-based models to 200 words in our designed system prompt as L-Defense. Since the explanation corresponding to the “half” prediction of L-Defense is derived from the combining of two competing justifications, its length is slighter longer than others.

### D.3 Ablation Study of Extraction Objective

As shown in Table 10, the results demonstrate the contribution of each objective in the extraction module.

Table 7: Hyper-parameters.

Hyperparm	RAWFC	LIAR-RAW
Competing Evidence Extraction		
Epoch	5	5
Batch Size	2	2
Learning Rate	1e-5	1e-5
$\gamma$ in Eq.(11)	0.9	0.5
Defense-based Inference		
Epoch	5	5
Batch Size	8	32
Learning Rate	1e-6	1e-6

Table 8: Explanations evaluation results of competing explanations using a 5-Point Likert scale rating by ChatGPT on RAWFC’s test set.

Gold veracity label	False		Half		True	
Given prior label	F	T	F	T	F	T
Informativeness	4.06	3.95	3.98	4.28	3.85	4.46
Soundness	4.21	3.88	4.09	4.10	3.92	4.45

Table 9: The average number of tokens per explanation generated by each method on the RAWFC and LIAR-RAW test sets.

Method	RAWFC	LIAR-RAW
Oracle	201.68	220.75
CofCED [40]	298.48	220.56
ChatGPT <sub>full</sub>	144.32	139.15
ChatGPT <sub>claim</sub>	128.71	150.97
L-Defense <sub>ChatGPT</sub>	266.61	225.52
L-Defense <sub>LLaMA2</sub>	305.50	175.38

Table 10: Ablation study on objective detailed in of veracity prediction on RAWFC.

Method	P	R	F1
L-Defense <sub>extractor</sub>	51.06	50.95	50.69
w/o KL-divergence obj	48.41	46.95	46.79
w/o classification obj	47.08	46.52	46.75

### D.4 More Case Studies

We provide more cases of the L-Defense’s predictions and explanations in Table 11. In general, the first two cases show that the informativeness and soundness of the truth side are at a higher level than the competing one, which proves our assumption. Moreover, we analyze error cases made by L-Defense on RAWFC’s test set, most of them fall into the failure due to the failure extraction of accurate evidence sentences, as the third case listed in Table 11.

## E FUTURE WORK

The advanced method to extract salient evidence can be further explored to provide accurate evidences and then gain improvements. Also, the advanced prompt method to improve the reasoning ability in fake news detection of LLM is valuable to explore. Furthermore, how to adaptively select suitable LLM for different claims can be investigated for further improvement.

**Table 11: More case studies. The “Gold” method denotes the gold label and explanations from Snopes. The two explanations generated by L-Defense are all listed for comparison.**

<b>Claim:</b> The U.S. Supreme Court has ruled against House Speaker Nancy Pelosi on impeachment.
<b>Method:</b> Gold, <b>Prediction:</b> false
<ul style="list-style-type: none"> <li>· A blog post that went viral in December 2019 created the false impression that the U.S. Supreme Court had issued a “landmark decision” favoring President Donald Trump in ongoing impeachment proceedings against him, thereby dealing a blow to House Speaker Nancy Pelosi. The post, “Supreme Court Ends Pelosi’s Reign of Terror with Landmark Decision on Impeachment,” was published on Dec. 17, 2019, and reported that “Nancy Pelosi just got spanked by the Supreme Court in a new impeachment ruling.” However, the story did not support what the headline proclaimed.</li> </ul>
<b>Method:</b> L-Defense <sub>ChatGPT</sub> , <b>Prediction:</b> false
<ul style="list-style-type: none"> <li>· What’s true: The claim is reasoned as True based on the following rationale: House Speaker Nancy Pelosi has been vocal in her opposition to President Trump’s Supreme Court nominee, Amy Coney Barrett. Pelosi referred to Barrett as an “illegitimate Supreme Court justice” during a news conference. This indicates a disagreement between Pelosi and the Supreme Court, suggesting that the Supreme Court has ruled against Pelosi’s position on impeachment. Additionally, Pelosi’s statement about not ruling out impeaching President Trump again to delay the Senate filling the Supreme Court vacancy implies that she is concerned about the Court’s influence on future impeachment proceedings. Given Pelosi’s opposition and concerns, it is reasonable to conclude that the U.S. Supreme Court has ruled against her on impeachment.</li> <li>· What’s false: The claim that the U.S. Supreme Court has ruled against House Speaker Nancy Pelosi on impeachment is false. There is no evidence or mention of such a ruling in the provided sentences. The sentences do mention various topics related to the Supreme Court, such as Joe Biden’s stance on court packing and the court’s upcoming hearing on the Trump administration’s challenge to Obamacare. However, none of these sentences support the claim that the Supreme Court has ruled against Pelosi on impeachment. Additionally, there is a mention of Pelosi’s comment on the Supreme Court’s ruling on financial records of the President, but this does not pertain to impeachment. Therefore, based on the lack of any supporting evidence or mention of such a ruling, the claim is deemed false.</li> </ul>
<b>Claim:</b> In February 2018, a pride of lions killed and ate a man suspected of being a poacher near Kruger National Park in South Africa.
<b>Method:</b> Gold, <b>Prediction:</b> true
<ul style="list-style-type: none"> <li>· Ingwēlala Private Nature Reserve in South Africa was shaken with the screams of a man who was attacked and killed by [a] Pride of Lion[s] in [a] private game reserve near the Kruger National Park. Before the lions were dispersed with the help of gunshots, they had eaten most of the body barring the head that was untouched.</li> </ul>
<b>Method:</b> L-Defense <sub>ChatGPT</sub> , <b>Prediction:</b> true
<ul style="list-style-type: none"> <li>· What’s true: The presence of sentences discussing illegal hunting and poaching activities in Kruger National Park, as well as the mention of a man being suspected of being a poacher, provide evidence that supports the claim. The sentence mentioning the man’s family seeking help from authorities after his death further strengthens the claim. Additionally, the sentence about the man and two others hunting illegally in the park and being surprised by an elephant suggests that the man’s death was a result of his illegal activities. The mention of previous incidents where suspected poachers have fallen victim to their prey in the park also supports the claim. Overall, the combination of these sentences provides a rationale for why the claim is reasoned as true.</li> <li>· What’s false: The claim is reasoned as False because there is no evidence to support the claim that a pride of lions killed and ate a man suspected of being a poacher near Kruger National Park in South Africa in February 2018. The sentences provided mention incidents of lion poaching and rhino poaching in the region, but there is no direct mention or evidence of a lion killing a suspected poacher. The sentences also discuss the Game Theft Act and the history of poaching in the area, but they do not provide any information about the specific incident mentioned in the claim. Additionally, there is mention of a lion skeleton being sold and the methods used by poachers, but again, no direct evidence of the claim. Therefore, based on the lack of supporting evidence, the claim is deemed False. )</li> </ul>
<b>Claim:</b> A Harvard study proved that Apple purposely slows down its older model iPhones to coincide with device releases and to boost sales of new models.
<b>Method:</b> Gold, <b>Prediction:</b> half
<ul style="list-style-type: none"> <li>· What’s true: Some recent iOS upgrades slow down the processing speeds of iPhones with degraded batteries.</li> <li>· What’s false: The processor slowdown is (according to Apple) intended to prevent unexpected power-related shutdowns and other problems in older iPhones that have batteries whose performance has degraded over time, not to boost sales of just-released newer iPhone models.</li> </ul>
<b>Method:</b> L-Defense <sub>ChatGPT</sub> , <b>Prediction:</b> false
<ul style="list-style-type: none"> <li>· What’s true: The presence of a study conducted by a Harvard University PhD student lends credibility to the claim that Apple purposely slows down its older model iPhones to coincide with device releases and boost sales of new models. This study supports long-standing suspicions among customers that the company engages in planned obsolescence to encourage upgrades. Additionally, the fact that Apple ships fewer iPhones in recent quarters suggests that the company may employ tactics to incentivize customers to purchase newer models. The negative reaction from consumers towards planned obsolescence, as well as the potential frustration for gamers who do not use Apple products, adds further weight to the claim. The study’s affiliation with Harvard University enhances its reliability, as the institution is recognized for its academic rigor. All these factors combined provide a streamlined rationale for the claim being true.</li> <li>· What’s false: The claim that Apple purposely slows down its older model iPhones to coincide with device releases and boost sales of new models is reasoned as False. A Harvard study is mentioned, but the sentences provided are unrelated or do not support the claim. The sentences mention a PHD student’s suspicion, a study on employee performance, the new Apple Watch’s health features, frustrations with new iPhone features, a description of the new camera feature, Tim Cook’s product announcement, complaints about slow performance after software updates, and a discussion of finding in a New York Times column. None of these sentences explicitly support the claim or provide evidence that Apple intentionally slows down older iPhones. Hence, the rationale for the claim being False is that there is no supporting evidence or relevant information to substantiate it.</li> </ul>