You Only Query Twice: Multimodal Rumor Detection via Evidential Evaluation from Dual Perspectives

Junyi Chen, Leyuan Liu*, Tian Lan, Fan Zhou, Xiaosong Zhang

University of Electronic Science and Technology of China

Abstract

Current rumor detectors exhibit limitations in fully exploiting responses to the source tweet as essential public opinions, and in explaining and indicating the reliability of the results obtained. Additionally, the joint utilization of both responses and the multimodal source content for detection presents challenges due to the heterogeneous nature of the data points. In this work, to address the first challenge, we initially prompt the Large Language Model (LLM) with both multimodal source content and the corresponding response set to extract contrasting evidence to enable maximal utilization of informative responses. To overcome the second challenge, we introduce an uncertainty-aware evidential evaluator to assess the evidence intensity from the multimodal source content and dual-sided reasoning, from which the final prediction is derived. As we model the second-order probability, we can effectively indicate the model's uncertainty (i.e., the reliability) of the results. The reasoning from the correct perspective also serves as a natural language-based explanation. To this end, the third challenge is also addressed as we fully leverage the available resources. Extensive experiments validate the effectiveness, uncertainty awareness in predictions, helpful explainability for human judgment, and superior efficiency of our approach compared to contemporary works utilizing LLMs.

1 Introduction

While social media platforms facilitate information exchange, they also enable rapid rumor dissemination. Existing research mainly combats this with content and response-based detection methods.

Content-based methods capitalize on data directly extracted from the source tweet. Initial research in this domain focused on learning textual representations of the source tweet (Zhang et al., 2015; Mikolov et al., 2013; Devlin et al., 2018). Subsequent studies have adopted multimodal learning approaches (Singhal et al., 2019; Wu et al., 2021; Qian et al., 2021), incorporating both image and text data, to discern more distinct patterns indicative of rumors. Conversely, response-based methods (Bian et al., 2020; He et al., 2021; Liu et al., 2023; Sun et al., 2022) utilize replies to the source tweet to gather social context, operating under the assumption that the nature of public reactions can provide signals helpful in distinguishing between different types of news.

Despite the partial effectiveness of these approaches, they exhibit several limitations: (1) Response-based methods may encounter challenges when informative evidence within certain responses is overshadowed by a preponderance of irrelevant replies or dominated by a unilateral, erroneous viewpoint. (2) Most existing methodologies lack robust mechanisms for explaining their results comprehensively and often fail to indicate the reliability of these results. (3) There is a paucity of studies that effectively integrate both visual and textual features from the source tweet alongside corresponding response characteristics to enhance detection accuracy.

To address these challenges, we introduce DEEP, a framework for rumor Detection via Evidential Evaluation from dual Perspectives. To overcome the first challenge, inspired by the powerful reasoning and extraction capabilities of Large Language Models (LLMs), we employ LLMs to reason both the truthfulness and falsehood of claims based on multimodal source tweets and their responses. This enables the maximal utilization of informative responses and the distillation of evidence for both supporting and opposing viewpoints, setting a comprehensive basis for evaluation. To address the second challenge, we deploy an uncertainty-aware evidential evaluator to assess the evidence intensity within the fused representation of multimodal

[*] Corresponding author

Proceedings of the 31st International Conference on Computational Linguistics, pages 3415–3427
January 19–24, 2025. ©2025 Association for Computational Linguistics

content and reasoning from dual perspectives to compute the prediction. This models the second-order probability, allowing for the effective computation of associated prediction uncertainty, thereby indicating the reliability of the results. The correct reasoning provided by the LLMs offers a sufficient explanation for the prediction behavior. To this end, the third challenge is addressed as we fully leverage the available multimodal content and corresponding responses. In summary, our contributions are delineated as follows:

- We introduce a dual perspectives reasoning module that effectively extracts both supporting and refuting evidence from multimodal sources and their responses.
- We develop an evidential evaluator that assesses the intensity of evidence based on the fused representation of multimodal content and side reasoning. This evaluator aids in reliably selecting the correct reasoning to elucidate the final prediction in an uncertainty-aware manner.
- We demonstrate that the proposed method leverages both multimodal content and its responses to achieve state-of-the-art performance in an uncertainty-aware manner. Empirical evidence also confirms that our method enhances human judgment and offers greater efficiency compared to contemporary works utilizing LLMs.

2 Problem Statement

Consider an instance (I_i, T_i, R_i) in a rumor detection dataset, where I_i represents the image of the source tweet content, T_i the text of the source tweet, and $R_i = [r_1, r_2, \ldots, r_n]$ the set of responses to the source tweet. Our objective is to develop a classification function that categorizes each instance as either a rumor or a non-rumor. For simplicity, subscripts may be omitted unless necessary for clarity.

3 Method

As depicted in Figure 1, our proposed framework comprises two modules: Dual Perspective Reasoner and Evidential Evaluator.

3.1 Dual Perspective Reasoner

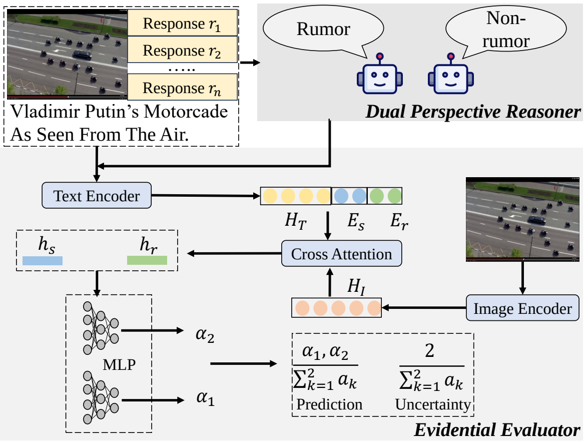


Figure 1: Schematic diagram of our proposed method.

Detecting tweets based on responses presents distinct challenges. Initially, online users often comment or reply casually to a source tweet, injecting chaos into the response set and obscuring explicit connections between informative responses and their relevance to the source tweet. Moreover, assessing the truthfulness of a source tweet may require indicative signals from responses that bring concrete supporting or refuting evidence. However, valid responses can become obscured by a mass of irrelevant replies or by a one-sided, incorrect opinion, as a recent study has shown (Wang et al., 2024).

On the other hand, LLMs are renowned for their superior reasoning capabilities when given specific input prompts. For instance, prompting an LLM with ’`You need to find concrete evidence from the multimodal source content and responses to support/refute the truthfulness of the source tweet`’ can potentially address the aforementioned challenges. Inspired by this capability and a recent study (Wang et al., 2024), our approach employs two LLMs to act as supporter and refuter of the source tweet’s truthfulness through role-playing. Taking the supporter role as an example, we use the following prompt:

> You are an evaluator tasked with reasoning why a source tweet is factual and non-misleading. Given a source tweet (I, T) and its corresponding response thread R, please find any possible evidence in both the source content and responses to support your claim that the source tweet is non-rumorous. If a response is used as evidence, cite its text. Your reasoning should be a con-

cise paragraph focused on the core points.

The prompt for the refuter contrasts this by seeking evidence to challenge the truthfulness of the source tweet. Engaging LLMs in this manner allows us to (1) distill the most helpful evidence from the responses, (2) ensure that the evidence is competing and comprehensive from both perspectives, thus maximizing the utilization of the responses and the source content and (3) remain the LLM attentive to details in the source content, such as cross-modal and logical consistency.

3.2 Evidential Evaluator

The role of the evaluator is to adaptively select the appropriate reasoning from the dual-sided arguments and render a final decision. Given that the reasoning provided by the LLM may still contain errors or noise, the evaluator must also assess the reliability of the final prediction and the chosen explanation, effectively quantifying the model's uncertainty. Motivated by recent advancements in deep uncertainty learning (Sensoy et al., 2020; Ulmer et al., 2021), we introduce an uncertainty-aware evaluator to provide conclusive results from an evidential perspective. The evaluator should first model the evidence intensity for each reasoning and derive the final prediction based on it. We begin by detailing how to fuse the multimodal content and respective reasoning for the final evaluation.

3.2.1 Multimodal Content and Reasoning Fusion

We first introduce how multimodal fusion is conducted in general. Given an image-text pair (I, T), we encode them separately as follows:

$$\boldsymbol{H}_T = TE(T), \quad \boldsymbol{H}_I = IE(I), \tag{1}$$

where TE and IE denote the text and vision-based transformers, producing text representation $\boldsymbol{H}_T = [\boldsymbol{h}_1^T, \ldots, \boldsymbol{h}_n^T] \in \mathbb{R}^{n\times d}$ and image representation $\boldsymbol{H}_I = [\boldsymbol{h}_1^I, \ldots, \boldsymbol{h}_m^I] \in \mathbb{R}^{m\times d}$. We first present a single-head cross-attention fusion:

$$CA(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{SOFTMAX}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

where $\boldsymbol{Q} = \boldsymbol{h}_1\boldsymbol{W}_Q, \boldsymbol{V}_h = \boldsymbol{h}_2\boldsymbol{W}_V, \boldsymbol{K} = \boldsymbol{h}_2\boldsymbol{W}_K$, and $\boldsymbol{W}_Q, \boldsymbol{W}_V, \boldsymbol{W}_K \in \mathbb{R}^{d\times d_k}$ are trainable weights. To achieve fine-grained aligned multimodal representation, we first compute the text-aligned image representation as follows:

$$\boldsymbol{h}^{T\to I} = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{CA(\boldsymbol{h}_i^T, \boldsymbol{h}_j^I, \boldsymbol{h}_j^I)}{m}, \tag{3}$$

where $\boldsymbol{h}^{T\to I} \in \mathbb{R}^{1\times d_k}$ is the text-aligned image representation. Similarly, we derive the image-aligned text representation $\boldsymbol{h}^{I\to T}$. The final multimodal content representation is obtained by a weighted sum of these two vectors:

$$\boldsymbol{h}_c = \boldsymbol{h}^{T\to I}\boldsymbol{W}_T + \boldsymbol{h}^{I\to T}\boldsymbol{W}_I, \tag{4}$$

where $\boldsymbol{W}_T$ and $\boldsymbol{W}_I$ are both trainable weights.

Now, let us denote the positive reasoning obtained from the reasoning module as E_s and the negative reasoning as E_r. These elements are individually concatenated with T (i.e., $T \oplus E_s$ and $T \oplus E_r$), serving as inputs for the text encoder as specified in Eq. 1. Subsequent operations, detailed from Eq. 3 to Eq. 4, are performed to derive the fused multimodal content and side reasoning representations, denoted as $\boldsymbol{h}_s$ and $\boldsymbol{h}_r$, respectively.

3.2.2 Evidential Learning

While one could ostensibly use a typical classifier with softmax function to obtain a prediction, such a naive implementation fails to accurately model real evidence intensity and associated uncertainty, as class probability derived by softmax only provides the first-order probability and a single-point estimation (Han et al., 2022). Besdies, softmax is notorious for producing over confident predictions.

Consequently, we turn to the concept of deep evidential learning for the classification task (Sensoy et al., 2018), which posits that the classification result $\hat{Y}$ is drawn from a variational Dirichlet distribution $Dir(\boldsymbol{\gamma}|\boldsymbol{\alpha})$, where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ represent the predicted probabilities and non-negative concentration parameters across two classes (i.e., rumor or non-rumor). The class probability can be computed as:

$$\hat{\boldsymbol{Y}} = \boldsymbol{\gamma} = \left[\frac{\alpha_1, \alpha_2}{\sum_{k=1}^{2}\alpha_k}\right] \tag{5}$$

Further, employing the principles of subjective logic (Jsang, 2018), we can model metrics related to evidence intensity and uncertainty:

$$b_k = \frac{\alpha_k - 1}{\sum_{k=1}^{2}\alpha_k}, \quad u = \frac{2}{\sum_{k=1}^{2}\alpha_k}, \tag{6}$$

where the belief mass b_k represents the evidence intensity for the k-th class, and uncertainty u is inversely correlated with it. The higher the focal evidence for a class, the lower the uncertainty. To jointly model evidence intensity and uncertainty from each viewpoint, the primary task is to output the concentration parameter $\boldsymbol{\alpha}$, achieved through the following procedure:

$$\begin{aligned} \alpha_1 &= MLP_s(\boldsymbol{h}_s) + 1, \\ \alpha_2 &= MLP_r(\boldsymbol{h}_r) + 1, \end{aligned} \tag{7}$$

where the Multi-layer Perceptron (MLP) is implemented with an activation function (e.g., ReLU) at the output layer. α_1 and α_2 represent the concentration parameters for the non-rumor and rumor classes, respectively, enabling effective computation of class probability via Eq. 5. Meanwhile, we can view α_1 as a proxy of truth evidence intensity while α_2 as the false one.

Maximize the Model Fit. Since we have derived the class probability distribution based on each viewpoint, we utilize a cross-entropy (CE) loss for them:

$$\mathcal{L}_{\text{Fit}} = CE(\hat{\boldsymbol{Y}}, \boldsymbol{Y}), \tag{8}$$

Minimize Evidence on Errors. Referring to Eq. 6, the task is to minimize the evidence (i.e., belief mass) on the incorrectly predicted class by pushing α_k to 1. Thus, we present the following loss:

$$\mathcal{L}_{\text{Err}} = KL\left(Dir(\boldsymbol{\gamma}|\hat{\boldsymbol{\alpha}}), Dir(\boldsymbol{\gamma}|\mathbf{1})\right) \tag{9}$$

where KL represents Kullback-Leibler divergence loss, and $\hat{\boldsymbol{\alpha}} = \boldsymbol{Y} + (\mathbf{1} - \boldsymbol{Y}) \odot \boldsymbol{\alpha}$ denotes the concentration parameters that are misleading for incorrect class predictions. $Dir(\boldsymbol{\gamma}, \mathbf{1})$ describes the uniform Dirichlet distribution where all concentration parameters are set to one. As each sample has only one possible ground truth, the evidence on incorrect reasoning is naturally minimized, avoiding high evidence on both classes. This also makes the uncertainty between correct and incorrect predictions more discriminative. We denote the final loss as:

$$\mathcal{L}_{\text{Eva}} = \lambda \mathcal{L}_{\text{Err}} + \mathcal{L}_{\text{Fit}} \tag{10}$$

where $\lambda \in [0, 1]$ is the annealing coefficient used to balance the two losses. This coefficient is gradually increased to allow the model sufficient exploration of the parameter space.

4 Experimental Evaluation

4.1 Experiment Setup

Dataset. We utilized two public datasets, PHEME (Zubiaga et al., 2017) and Twitter (Ma et al., 2017), for our experiments. Each instance in these datasets includes a multimodal source tweet, its associated responses, and a ground truth label indicating whether it is classified as a rumor or not. Note that the Twitter dataset is the combined version of two tiny datasets Twitter15 and Twitter16. Detailed pre-process steps and data statistics are provided in Appendix A.1.

Baselines. We selected a collection of state-of-the-art baselines categorized into several groups: (C1) Content-based methods including BERT (Devlin et al., 2018), EANN (Wang et al., 2018), SAFE (Zhou et al., 2020), and KDCN (Sun et al., 2023); (C2) Response-based methods including RDEA (He et al., 2021), TrustRD (Liu et al., 2023), GACL (Sun et al., 2022), and MFAN (Zheng et al., 2022); and (C3) LLM-based methods including GenFEND (Nan et al., 2024), DELL (Wan et al., 2024), and L-Defense (Wang et al., 2024). Detailed baseline descriptions are provided in Appendix A.2. These LLM-based methods primarily utilize textual modality from source content and responses. To ensure a fair comparison of input modalities, we substituted the original LLM used in their framework with a vision LLM, enabling the handling of multimodal source tweets to collect the output. Note that we also experimented with textual versions of the LLM-based methods and the performance difference is minor, with the visual version sometimes performing slightly better. Further details are provided in Appendix A.4.

Evaluation Protocol. The datasets were split in a ratio of 6:1:3 for training, validation, and testing, respectively. All methods implemented an early stop mechanism after 20 epochs based on performance on the validation set. The best checkpoint from the validation set was then used on the test set to obtain final results. Following previous studies, we report metrics including Accuracy, macro Precision, Recall, and F1. Note that we have also implemented a specific measure for all LLM-based methods (including ours) to avoid potential data leakage. For implementation details of our method and baselines, please refer to Appendix A.3.

Model	Twitter				PHEME			
	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1
BERT	0.769	0.697	0.765	0.711	0.810	0.806	0.792	0.799
EANN	0.834	0.763	0.835	0.785	0.834	0.809	0.822	0.815
SAFE	0.838	0.765	0.844	0.792	0.832	0.801	0.819	0.810
KDCN	0.851	0.781	0.853	0.805	0.845	0.832	0.826	0.829
RDEA	0.859	0.789	0.856	0.812	0.833	0.815	0.818	0.816
TrustRD	0.875	0.808	0.881	0.833	0.857	0.841	0.830	0.835
GACL	0.870	0.806	0.882	0.832	0.860	0.852	0.829	0.841
MFAN	0.881	0.830	**0.906**	0.856	0.865	0.857	0.842	0.849
GenFEND	0.890	0.851	0.880	0.865	0.865	0.842	0.852	0.848
DELL	0.887	0.841	0.876	0.858	0.855	0.850	0.822	0.843
L-Defense	0.892	0.860	0.890	0.872	0.861	**0.871**	0.828	0.849
DEEP	**0.916**	**0.872**	0.903	**0.885**	**0.889**	0.868	**0.879**	**0.872**

Table 1: Performance comparison. The best metric is in **bold** and the runner-up is underlined

4.2 Performance Comparison

The performance of all methods is reported in Table 1, from which several insights can be drawn. (1) Content-based methods that focus solely on textual modality (i.e., BERT), exhibit the poorest performance. This outcome is anticipated, as tweet content is typically short and informal, providing limited textual cues for effectively distinguishing between rumors and non-rumors. Conversely, the inclusion of image representation as additional heterogeneous data points significantly enhances performance. (2) Response-based methods generally outperform content-based approaches. Unlike methods that concentrate solely on the source content, response-based methods leverage additional indicative signals from public interactions. The response graphs established by them can also aid in discriminating rumors. Notably, MFAN, which also incorporates image modality, achieves superior results compared to other response-based methods. (3) LLM-based methods establish a new benchmark in rumor detection. Specifically, DELL leverages LLMs to generate responses to source tweets through role-playing and defines multiple proxy tasks to obtain aggregated predictions, yielding promising results. GenFEND follows a similar direction to DELL but also integrates actual responses with generated ones. Meanwhile, L-Defense employs a pre-trained extraction module to obtain competing evidence for LLM reasoning, achieving the best result among the baselines. However, L-Defense may underperform if the extractor fails to distill valid evidence from the responses, and it lacks fine-grained quantification of evidence intensity from both sides to derive uncertainty-aware predictions. In contrast, our approach utilizes LLMs to directly extract evidence from responses and perform reasoning, which obviates the need for pre-training an extractor module and consistently identifies informative evidence within the sea of responses. Furthermore, our method quantifies evidence intensity from both perspectives and derive the prediction in an uncertainty-aware manner, enhancing accuracy and trustworthiness.

4.3 Ablative Study

To evaluate our principal design motivations, we have developed the following variants: (A1) LLM utilization strategy: w/o Supporter, which exclusively employs the LLM for negative reasoning; w/o Refuter, which uses the LLM solely for positive reasoning; and w/ Direct, where the LLM is directly fed both the source tweet and responses to assess veracity. (A2) Evaluator strategy: w/ Concat, which concatenates the fused multimodal positive and negative reasonings and inputs them into a softmax-based classifier; w/ Attention, where a typical attention mechanism is implemented to integrate the fused multimodal dual reasoning before input to the classifier. (A3) Input modality: w/o Visual, which only leverages the textual modality as input to the framework.

The results, depicted in Figure 2, allow us to extract several insights. Firstly, improper utilization of LLMs can lead to a notable degradation in performance. Specifically, prompting the LLM from a single perspective may result in a biased

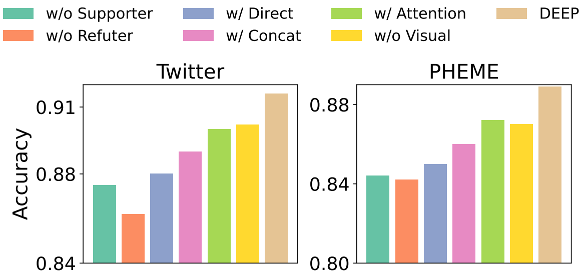
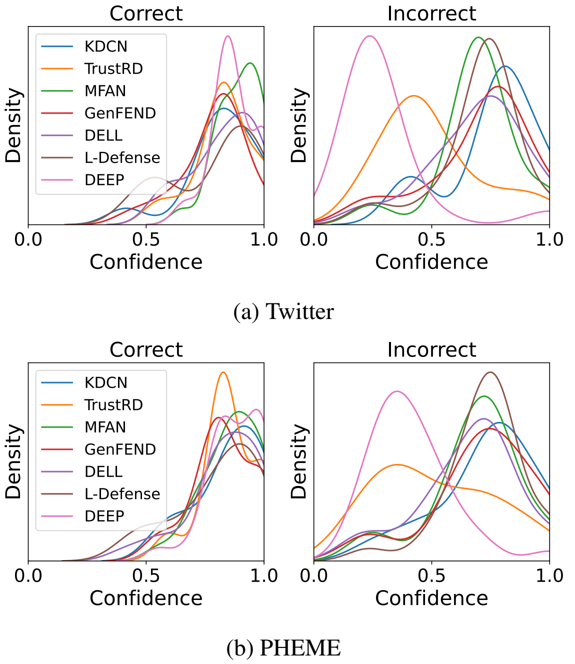


Figure 2: Ablative study on model variants.

evaluation, potentially confusing the final evaluator. Moreover, directly prompting the LLM with source tweets and responses may cause the model to overlook counterarguments, which are crucial for accurate rumor detection. Secondly, failing to model the evidence intensity from each viewpoint and deriving the final result based on associated uncertainty results in suboptimal performance. In particular, merely concatenating reasonings to derive a class probability distribution provides only a coarse evaluation of evidence soundness without fine-grained quantification. Similarly, while an attention-based mechanism mitigates this issue by focusing on potentially more informative reasoning, it still falls short in dynamically selecting more reliable and coherent one, as the attention weights cannot be equated with evidence intensity or model confidence. Finally, integrating the visual modality to conduct multimodal rumor detection is beneficial for performance.

4.4 Uncertainty Analysis

A significant advantage of our proposed method is its awareness of uncertainty in predictions. Specifically, the method is designed to exhibit lower uncertainty for correct predictions and higher uncertainty for incorrect ones. To validate this characteristic, we visualized the sampled uncertainty density distribution of both correct and incorrect predictions and compared these with selected baselines. To the best of our knowledge, TrustRD is the only baseline equipped with prediction uncertainty estimation, employing Bayesian variational inference. In contrast, other baselines predominantly use a softmax-based classifier to derive the probability of predictions. For a fair comparison, we normalized the final class probabilities of these models within the interval [0,1] as the model confidence. In other words, the higher the confidence, the lower the uncertainty and vice versa.



(a) Twitter

(b) PHEME

Figure 3: Uncertainty analysis with respect to correct and incorrect classification.

The results, illustrated in Figure 3, yield several insights. Firstly, most baselines fail to provide trustworthy confidence levels regarding correct and incorrect predictions. This suggests that modeling the first-order probability as the final prediction may not be reliable and could potentially mislead decision-makers with erroneous predictions. TrustRD, the best uncertainty-aware baseline, can differentiate the uncertainty distribution between correct and incorrect predictions to some extent. However, its approach requires computing 10 iterations per sample and relies on parameter adversarial learning, both of which are computationally intensive and not as accurate as our method. Secondly, Our proposed method adeptly differentiates between the uncertainty distributions of correct and incorrect predictions, requiring only a single computation per sample. This is attributed to modeling fine-grained evidence intensity to quantify the prediction uncertainty and effective uncertainty discriminability. Consequently, compared to existing methods based on LLMs aiming to produce explanations aiding human decisions, our model offers an additional advantage: it can reliably indicate the trustworthiness of the obtained explanations.

4.5 LLM-based Methods Comparison

4.5.1 Helpfulness Evaluation

A principal advantage of LLM-based methods is their ability to generate natural language-based ex-

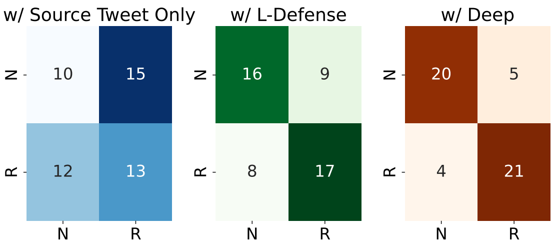


Figure 4: Helpfulness evaluation: Human judgment of rumors based on source tweets alone versus outputs from LLM-based methods.'N' denotes 'non-rumor' and 'R' denotes 'rumor'.

planations, facilitating human involvement in the decision process. Given that L-Defense is the only baseline designed to produce explanations to aid human decisions, we included it alongside our proposed method in a human evaluation study. Specifically, we enlisted 10 linguistic experts to assess the truthfulness of 50 source tweets using the output from LLM-based methods. For L-Defense, we provided the prediction and its corresponding explanation, while for our method, we additionally provided the prediction uncertainty.

The results, depicted in Figure 4, allow us to extract several insights. (1) Relying solely on the source tweet resulted in the poorest performance among annotators, as it demands extensive topical knowledge, which is often unattainable for each sample. (2) The inclusion of outputs from LLM-based methods significantly enhanced performance. This improvement confirms that appropriately integrating LLMs within the rumor detection pipeline can increase the explainability of the task, thereby guiding humans to more accurate decisions. (3) The output from our proposed method enabled annotators to achieve superior performance compared to that under L-Defense. This outcome can be attributed to two factors. First, our method more effectively extracts valuable clues from the input than L-Defense, whose pre-trained extractor may occasionally overlook existing key evidence. Second, due to the modeling of fine-grained evidence intensity, our prediction is more accurate and the uncertainty indicator provides an additional layer of verification, informing annotators about the reliability of the predictions and explanations.

4.5.2 Efficiency Analysis

A significant concern with LLM-based methods is their requirement for additional queries to collect data from LLMs during both training and inference phases. To address this, we conducted an efficiency analysis involving four LLM-based methods to determine whether their accuracy improvements justify the additional computational intensity. This analysis encompassed two aspects: training efficiency (i.e., the time required to complete training) and inference efficiency (i.e., the time required for end-to-end inference per sample).

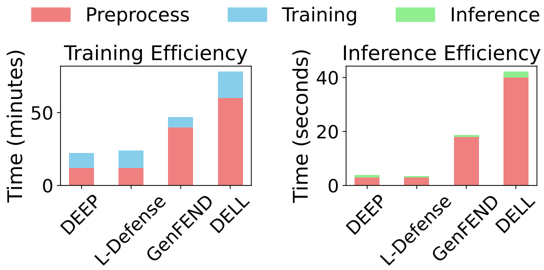


Figure 5: Efficiency analysis of LLM-based methods on the Twitter dataset. 'Preprocess' refers to the time needed to query the LLM. For a fair comparison, all experiments were conducted on a single A100 GPU using frozen LLaVA-13B.

The results, illustrated in Figure 5, indicate that GenFEND and DELL demand extensive time for both training and inference. Specifically, GenFEND necessitates querying the LLM 30 times per sample, while DELL requires even more extensive iterations to construct the response graph. In contrast, our proposed method and L-Defense demonstrate greater and comparable efficiency, requiring minimal queries to the LLM.

4.6 Early Rumor Detection

We incorporated our proposed method and selected baselines to conduct an early rumor detection experiment, examining various proportions of actual responses in chronological order. Methods that are not sensitive to variations in response quantity were excluded.

The results, presented in Figure 6, indicate that most response-based methods are highly sensitive to the quantity of responses, with limited responses in early detection scenarios significantly constraining their performance. In contrast, LLM-based methods, particularly our proposed method, demonstrate robustness in scenarios with scant responses. On the one hand, this resilience can be attributed to the impressive reasoning capabilities of LLMs that effectively utilize even minimal content and insufficient responses. On the other hand, the results validate that our dual-sided guidance approach enables LLMs to effectively extract contrasting evidence from multimodal content and helpful responses

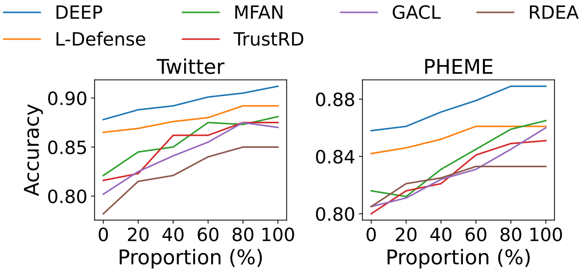


Figure 6: Early rumor detection comparison.

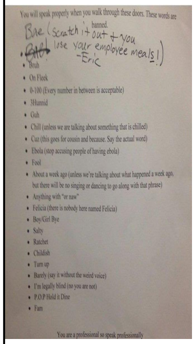
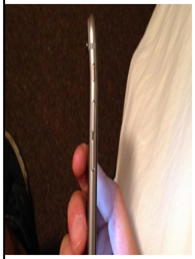
 You will speak properly when you walk through these doors. These words are banned. Bae (Scratch it out + you lose your employee meals!) -Eric • ~~Bae~~ • Bruh • On Fleek • 0-100 (Every number in between is acceptable) • 3Hunnid • Guh • Chill (unless we are talking about something that is chilled) • Cuz (this goes for cousin and because. Say the actual word) • Ebola (stop accusing people of having ebola) • Fool • About a week ago (unless we're talking about what happened a week ago, but there will be no singing or dancing to go along with that phrase) • Anything with "or naw" • Felicia (there is nobody here named Felicia) • Boy/Girl Bye • Salty • Ratchet • Childish • Turn up • Barely (say it without the weird voice) • I'm legally blind (no you are not) • P.O.P Hold it Dine • Fam You are a professional so speak professionally **Text**: No baes allowed: a Chick-fil-A manager has banned 22 hip words	**Explanation:** This is likely a rumor based on several points of skepticism expressed in the response thread and the source tweet context. Multiple replies indicate doubt or confusion about the validity of the claim, with one user explicitly stating, **"prob fake, still funny doe,"** suggesting they believe the story isn't true but find it amusing nonetheless. Additionally, another user sarcastically remarks, **"Crazy! Can't believe a business expects its employees to act in a professional manner- outrageous!",** implying that the premise of banning slang might be exaggerated or fabricated for effect. Moreover, **the general tone of the responses, filled with humor and disbelief,** further supports the idea that the original tweet is not being taken seriously by many readers. Besides, **the provided image seems not a formal document from Chick-fil-A. (0.12)**
 **Text**: iPhone 6 Plus users report the body is bending in pants pockets	**Explanation:** The truthfulness of the tweet is supported by both the image provided and the responses in the thread. **The image clearly shows an iPhone 6 Plus with a noticeable bend**, directly validating the claim made in the tweet. Responses such as **"@smh roflmao.and how much did this cost?cooking foil over some plakky"** and **"@smh Just like wondering why your glasses broke cos they were in your hip pocket when you sat down??? Logical!"** indicate that users are acknowledging and discussing the bending issue, with some even humorously commenting on the situation. The combination of visual evidence and user reactions both supports the truthfulness. **(0.86)**

Figure 7: Two real-world rumor cases: The upper case is correctly identified, and the lower case is misclassified. Key evidence is highlighted, with the model's uncertainty indicated in parentheses.

so far, even in the absence of substantial public interaction. Consequently, our proposed method represents a generalizable framework capable of adapting to varying response availability in rumor detection tasks.

4.7 Case Study

To further substantiate the efficacy of our model, we present two case studies involving real-world rumors analyzed using our model's outputs, explanations, and indicated uncertainties. In these cases, one rumor is accurately identified, while the other is erroneously classified as a non-rumor.

The cases are illustrated in Figure 7. For the accurately identified rumor, the model effectively garners concrete refuting evidence from the responses, encapsulating the semantic essence of the response thread at a high level. Additionally, the model identifies suspicious elements within the provided image, collectively supporting the conclusion that the source tweet disseminates misleading information. Conversely, for the misclassified sample, the evidence derived from the responses is weak and includes erroneous reasoning (the model misinterprets challenging response semantics as supportive). We hypothesize that the misclassification arises due to the absence of strong counterarguments, both from individual responses and at the semantic level of the response thread, coupled with strong cross-modal consistency in the multimodal source tweet. On the other hand, the model exhibits high uncertainty regarding this prediction (as there also lacks strong supportive semantics from the responses), a factor that could also be informative.

5 Related Work

Initial recognized efforts in rumor detection employed content-based methods, using neural networks like BERT (Devlin et al., 2018), Word2Vec (Mikolov et al., 2013), and TextCNN (Zhang and Wallace, 2015) to capture text representations of source content. Recognizing the limitations of short, informal tweet text, later approaches integrated visual features to develop multimodal representations, based on the assumption that consistency between text and images suggests authenticity (Sun et al., 2023). Recently, response-based methods have gained prominence, utilizing response threads as public wisdom. These methods often create bi-directional propagation graphs (Bian et al., 2020) and apply data augmentation (He et al., 2021) to enhance representation expressiveness while addressing noise (Sun et al., 2022; Liu et al., 2023) to improve robustness. However, most focus solely on textual responses due to challenges in integrating disparate visual data, a gap partially addressed by models like MFAN (Zheng et al., 2022) through hierarchical attention mechanisms. The rise of LLMs has further diversified approaches. Methods like GenFEND (Nan et al., 2024) and DELL (Wan et al., 2024) employ LLMs for role-playing simulations to enhance detection, while L-Defense (Wang et al., 2024) extracts evidence from responses for defense-based reasoning. Our work extends these concepts by using LLMs directly as evidence extractors, avoiding additional training phases and is more effective at extracting valid evidence. We also uniquely model dual-sided reasoning evidence intensity, enabling uncertainty-aware predictions that enhance reliability and explainability, and focus on multimodal rumor detection.

6 Conclusion

In this work, we introduce DEEP, a multimodal rumor detection framework that leverages the LLM to analyze content from dual perspectives and incorporates a parameterized evidential evaluator to assess contrasting evidence. We highlight several key findings: (1) the LLM effectively extracts important clues from a set of responses, and inspiring it to consider both positive and negative perspectives ensures comprehensive clues and evidence; (2) evidential learning proves effective and efficient in assessing noise and uncertainty from each side, adaptively selecting the more reliable one; and (3) as a combination of the above, our method produces more trustworthy and explainable predictions, aiding manual rumor evaluation in real-world contexts.

Limitation

While our work has achieved state-of-the-art performance and introduced several notable advantages compared to previous and contemporary studies, it can still be improved in the following directions: (1) Although we have empirically demonstrated that our proposed method is significantly more efficient compared to other contemporary work utilizing LLMs, querying the LLM still incurs a non-negligible computational cost compared to traditional deep neural networks. To address this, we will explore more efficient versions (e.g., via knowledge distillation) with minimal performance sacrifice. We also anticipate natural advancements in LLM efficiency over time. Additionally, we have utilized the well-known open-source LLM, LLaVA, but have not experimented with the well-known closed-source LLM, GPT-4, due to potential extensive API financial costs. Nevertheless, our proposed method is a general paradigm that can be applied to both open-source and closed-source LLMs. Given that GPT-4 is known as the most powerful LLM currently available, applying our work to GPT-4 could foreseeably achieve even better performance. (2) Our method faces challenges when the number of responses per instance significantly increases, particularly regarding the LLM's ability to extract the most informative ones. While advancements in LLM window sizes can partially address this issue, they raise concerns about efficiency. In this context, (Wang et al., 2024) provides insights by using a pre-trained neural network to extract seed responses, though it may occasionally miss important ones. Future work could explore more effective extraction modules based on related studies in RAG. (3) Due to the constraints of the selected dataset, our focus was limited to binary classification tasks (i.e., rumor or non-rumor). Future work could explore more fine-grained classification levels (e.g., half-true rumors). In this regard, multi-view learning could be effective, where fine-grained prediction logits are derived from each view, and a fusion module combines them to produce the final result.

Ethical Consideration

The spread of rumors on social media is a significant societal issue, which serves as the primary motivation for our proposed method. However, since our method prompts the LLM to consider dual perspectives, it reasons why a piece of news could be both a rumor and not a rumor simultaneously. This feature could potentially be exploited by malicious users to create seemingly reasonable defenses for actual rumors, or to deceive our system and other real-world rumor detectors. Additionally, the data used in our experiments, which includes responses from online users, could raise sensitive privacy concerns if not appropriately anonymized. To address these issues, we commit to ensuring responsible access to our resources, including both the data and the code used to replicate our framework. Furthermore, all released preprocessed data will be anonymized to protect sensitive information about online users, in accordance with established Twitter policies.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No. U2336204), the Sichuan Province Sci-Tech Plan (Grant No. 2024YFG0005), and the National Natural Science Foundation of China (Grant No. 62272089).

References

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 549–556.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. 2022. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):2551–2566.

Zhenyu He, Ce Li, Fan Zhou, and Yi Yang. 2021. Rumor detection on social media with event augmentations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2020–2024.

Audun Jsang. 2018. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated.

Leyuan Liu, Junyi Chen, Zhangtao Cheng, Wenxin Tai, and Fan Zhou. 2023. Towards trustworthy rumor detection with interpretable graph structural learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4089–4093.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. Let silence speak: Enhancing fake news detection with generated comments from large language models. *arXiv preprint arXiv:2405.16631*.

Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 153–162.

Murat Sensoy, Lance Kaplan, Federico Cerutti, and Maryam Saleki. 2020. Uncertainty-aware deep classifiers using generative models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5620–5627.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE.

Mengzhu Sun, Xi Zhang, Jianqiang Ma, Sihong Xie, Yazheng Liu, and S Yu Philip. 2023. Inconsistent matters: A knowledge-guided dual-consistency network for multi-modal rumor detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12736–12749.

Tiening Sun, Zhong Qian, Sujun Dong, Peifeng Li, and Qiaoming Zhu. 2022. Rumor detection on social media with graph adversarial contrastive learning. In *Proceedings of the ACM Web Conference 2022*, pages 2789–2797.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. 2021. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *arXiv preprint arXiv:2110.03051*.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. Dell: Generating reactions and explanations for llm-based misinformation detection. *arXiv preprint arXiv:2402.10426*.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM on Web Conference 2024*, pages 2452–2463.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857.

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 2560–2569.

Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

Zili Zhang, Ziqiong Zhang, and Hengyun Li. 2015. Predictors of the authenticity of internet health rumours. *Health Information & Libraries Journal*, 32(3):195–205.

Jiaqi Zheng, Xi Zhang, Sanchuan Guo, Quan Wang, Wenyu Zang, and Yongdong Zhang. 2022. Mfan: Multi-modal feature-enhanced attention networks for rumor detection. In *IJCAI*, volume 2022, pages 2413–2419.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. : Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on knowledge discovery and data mining*, pages 354–367. Springer.

Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2017. Exploiting context for rumour detection in social media. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I 9*, pages 109–123. Springer.

Statistic	Twitter	PHEME
# Source Tweet	1642	2018
# Avg. Response	35	15
# Rumor	506	590
# Non-rumor	1136	1428
# Image	1056	2018

Table 2: Dataset statistics of Twitter and PHEME

A Appendix

A.1 Data Statistics

We have amalgamated the Twitter15 and Twitter16 datasets into a unified Twitter dataset, from which duplicate entries have been removed. We have consolidated the labels "true rumor" and "non-rumor" into a single "non-rumor" label, while retaining "false-rumor" as the "rumor" label. For PHEME, we use the publicly available dataset. Additionally, responses posted after the official debunking deadlines were excluded to prevent potential data leakage. Detailed data statistics are provided in Table 2.

Motivation for Adopting a Binary Setting for the Twitter Dataset. While the original Twitter dataset is a four-class prediction task, we adopt a binary classification setting in this study for greater real-world practicability and utility. Specifically, according to the original documentation (Ma et al., 2017), the "unverified" label refers to events unverified at the time the dataset was constructed. However, based on our manual inspections, many of these can now be verified, leading to potential label inconsistencies in the original setting that could mislead the model. Besides, as both the original true-rumor and non-rumor labels represent truthful events, we focus on a binary classification task distinguishing rumors from non-rumors.

A.2 Baseline Descriptions

- BERT (Devlin et al., 2018): A pre-trained language model utilizing deep bidirectional transformers. We employ BERT to derive the textual representation of posts for classification purposes.

- EANN (Wang et al., 2018): Employs an event adversarial neural network to extract event-invariant features from images and texts for rumor detection.

- KDCN (Sun et al., 2023): A knowledge-based multimodal rumor detection model that learns dual consistency between image-text content and entity-common knowledge correlations.

- RDEA (He et al., 2021): A response-based method that uses graph augmentation strategies to learn robust propagation representations for each selected instance.
- TrustRD (Liu et al., 2023): A trustworthy rumor detector that initially eliminates noise within rumor propagation and then applies Bayesian variational inference to quantify prediction uncertainty.
- GACL (Sun et al., 2022): A response-based rumor detector using adversarial contrastive learning to generate adversarial samples, thereby enhancing the model's discriminative capabilities.
- MFAN (Zheng et al., 2022): A multimodal rumor detector that utilizes both the multimodal content and corresponding responses, employing a hierarchy attention-based mechanism to fuse different modalities for prediction.
- GenFEND (Nan et al., 2024): An LLM-based rumor detector that uses the LLM to generate pseudo responses from different user groups to uncover insights from users unlikely to comment in real-world scenarios.
- DELL (Wan et al., 2024): An LLM-based approach that simulates user interactions with various types of source tweets and defines a set of proxy tasks for the LLM to solve, thereby enhancing rumor detection.
- L-Defense (Wang et al., 2024): An LLM-based method that initially extracts competing evidence from responses using a pre-trained extractor and then prompts the LLM to reason about the dual-sided veracity of the given source tweet.

A.3 Implementation Details

We developed our framework using Pytorch. For the LLM, we selected `LLaVA-13B`. The text encoder implemented was the pre-trained `bert-uncased-english`, and for image encoding, we employed the state-of-the-art vision Transformer `google/vit-base-patch16-224`. The hidden dimensions d and d_k were set to 768. All MLPs are two-layered with activation function ReLu. We configured the learning rate to 1×10^{-4}, set the batch size to 16, and utilized Adam as our optimizer for training. Note that the LLM is frozen.

For the implementation of baselines, we reproduced their work using the officially released code and fine-tuned the parameters to achieve optimal performance.

Measures on Data Leakage for LLM-based Methods. Considering that most LLMs are pre-trained with data available beyond the deadlines of the datasets used, to prevent potential data leakage (i.e., the LLM might already know the ground truth of the instances), we implemented a specific measure. We prompted the LLM with the following query for each dataset instance:

> Based on your own factual knowledge, do you know the ground truth label of the tweet $[s, u, t]$ awaiting rumor verification? Please select only one of the following options: rumor, non-rumor, or I don't know.

The responses from the LLM were then subjected to a filtration process. If the LLM provided a correct answer with a coherent explanation, it indicated that the model was already aware of the truth without our designed prompts and workflow, which would be considered data leakage.

A.4 Additional Performance Comparison on Baseline Variants

In our pilot study, We first use the original textual LLM settings for baselines as described in their papers and then experiment by replacing the textual LLM with a visual one in our main experiments. Here, we present the performance of LLM baselines using textual LLMs.

As shown in Table 3, the performance of all LLM baselines using textual versions is consistent with that of visual LLMs under our main experimental settings, with visual LLMs sometimes performing better. Therefore, for fair comparison, we use visual LLMs for all LLM baselines. Another thing needs to be noted is that the trainable parameter size of our method is comparable to the baselines, approximately 300M.

Model	Twitter		PHEME	
	Acc	F1	Acc	F1
GenFEND	0.886	0.859	0.859	0.842
DELL	0.889	0.855	0.851	0.839
L-DEFEND	0.895	0.871	0.859	0.848

Table 3: Additional experiments on baseline variants by using textual LLMs. The estimated trainable parameter size for baselines is approximately 300-400M.