From Predictions to Analyses: Rationale-Augmented Fake News Detection with Large Vision-Language Models

Xiaofan Zheng
Xi'an Jiaotong University
Xi'an, Shaanxi, China
zxf_xjtu@stu.xjtu.edu.cn

Zinan Zeng
Xi'an Jiaotong University
Xi'an, Shaanxi, China
2194214554@stu.xjtu.edu.cn

Heng Wang
Xi'an Jiaotong University
Xi'an, Shaanxi, China
wh2213210554@stu.xjtu.edu.cn

Yuyang Bai
Xi'an Jiaotong University
Xi'an, Shaanxi, China
1206944633@stu.xjtu.edu.cn

Yuhan Liu
Xi'an Jiaotong University
Xi'an, Shaanxi, China
lyh6560@stu.xjtu.edu.cn

Minnan Luo*
Xi'an Jiaotong University
Xi'an, Shaanxi, China
minnluo@xjtu.edu.cn

Abstract

The rapid development of social media has led to a surge of eye-catching fake news on the Internet, with multimodal news comprising both images and text being particularly prevalent. To address the challenges of Multimodal Fake News Detection (MFND), numerous supervised task-specific Multimodal Small Language Models (MSLMs) have been developed. However, these models lack the breadth of knowledge and the depth of language understanding, which results in unsatisfactory adaptability, generalization, and explainability performance. To address these issues, we attempt to introduce Large Vision-Language Models (LVLMs), aiming to leverage the common sense understanding and logical reasoning abilities of LVLMs for the MFND task. We observed that LVLMs can generate reasonable analyses of news content from specific angles. However, when it comes to synthesizing these analyses for final judgment, their performance declines significantly, failing to meet the accuracy benchmarks set by existing MSLMs detection models. This reflects the need for a more effective way for LVLMs, which have not undergone task-specific training, to utilize their knowledge and capabilities. Based on these findings, we propose the **E**xplainable **A**daptive **R**ationale-**A**ugmented **M**ultimodal (EARAM) framework, which adaptively uses MSLMs to extract useful rationales from the multi-perspective analyses of LVLMs. After making judgments based on these rationales, EARAM then assists LVLMs in generating more reliable explanations. Extensive experiments demonstrate that our model not only achieves state-of-the-art results on widely used datasets but also significantly outperforms other models in terms of generalization and explainability.

CCS Concepts

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

Keywords

Fake News Detection; Large Vision-Language Models; Explainable

ACM Reference Format:
Xiaofan Zheng, Zinan Zeng, Heng Wang, Yuyang Bai, Yuhan Liu, and Minnan Luo*. 2025. From Predictions to Analyses: Rationale-Augmented Fake News Detection with Large Vision-Language Models. In *Proceedings of the ACM Web Conference 2025 (WWW '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3696410.3714532

1 Introduction

In the contemporary information landscape, fake news propagates with unprecedented velocity and breadth, posing significant threats to social stability [1], public safety [36], and individual rights [41].

With the proliferation of social media platforms [39], we have observed that fake news exhibits distinct multimodal characteristics, containing both text and image information [35]. Multimodal fake news enhances its appeal by combining text and images, making it more deceptive and influential than purely text-based ones. Multimodal Fake News Detection (MFND), which involves comprehensive analyses of textual content, image features, and other external knowledge, offers a more robust approach to capturing the nuanced characteristics of fake news. Consequently, the development of more accurate and practical MFND techniques has become increasingly critical [34].

However, most existing fake news detection techniques are based on supervised task-specific Multimodal Small Language Models (MSLMs) [11]. Due to the limitation of the training datasets and the inherent constraints of the MSLMs, these models often lack the sufficiently strong ability to make common-sense judgments, deeply understand real-world contexts, and analyze the intrinsic logic between images and text [43]. For example, they only recognize superficial signals from images and text, and oversimplify the problem to merely check the semantic consistency between modalities, ignoring the deeper logic of whether the images and text can corroborate each other. **So how do these limitations hinder the effectiveness of such methods?** Experiments and research have shown that although these MSLMs have achieved good results on their respective datasets, due to the limitations in knowledge and constrained logical reasoning ability, most methods exhibit significant performance drops when encountering news topics or entities not covered in the training data [10, 28]. But real-world news is

*Corresponding author: Minnan Luo, School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China.

WWW '25, Sydney, NSW, Australia
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1274-6/25/04
https://doi.org/10.1145/3696410.3714532

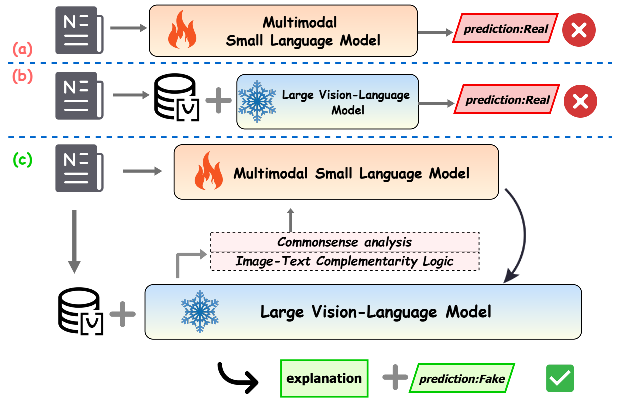

(a)
Multimodal Small Language Model
prediction:Real
(b)
Large Vision-Language Model
prediction:Real
(c)
Multimodal Small Language Model
Commonsense analysis
Image-Text Complementarity Logic
Large Vision-Language Model
explanation
prediction:Fake

Figure 1: Illustration of the roles of LVLMs and MSLMs.(a) Using only MSLMs lacks common sense and logical reasoning abilities. (b) Using only LVLMs can not integrate and synthesize decisions in specific domains. (c) Conducting multi-perspective analysis through LVLMs, then utilizing MSLMs to integrate the analysis and make judgments, and finally assisting LVLMs in generating explanations.

characterized by considerable dynamism and diversity. News topics and entities continuously evolve and expand over time and across different reporting platforms [63]. This severely impacts the generalization ability of MSLMs, making it difficult for them to be effectively applied in real-world applications. At the same time, these models can only make judgments and are unable to provide users with explanations for their judgment results [43].

Recent advancements in large vision-language models (LVLMs) such as GPT-4o [30], LLava [17], and InternVL2 [4] have demonstrated remarkable capabilities in cross-modal reasoning and real-world context understanding [52, 58]. Trained on vast text-image datasets, these models encompass extensive global knowledge and align closely with human preferences, achieving impressive results across various downstream tasks [53]. Despite these achievements, the potential of LVLMs in the domain of fake news detection remains largely unexplored [51, 57, 61]. To evaluate the capabilities of LVLMs in this domain, we conduct a series of experiments utilizing two prominent models: LLaVA-Next and InternVL2. We implement advanced prompt engineering techniques, including few-shot learning and chain-of-thought reasoning [49], as well as Retrieval-Augmented Generation (RAG) to mitigate potential hallucinations [25] in LVLMs. Our experiments and previous research consistently indicate that while LVLMs can provide reasonable analyses from specific perspectives, their direct performance in making final judgments falls short of typical MSLMs [9, 29, 33]. We attribute this discrepancy to potential differences between the training corpora of LVLMs and the specific task requirements, as well as the primary training objectives of LVLMs, which focus on understanding and generating multimodal content rather than integrating analyses and making precise classifications within specific narrow domains [58].

To address these challenges and leverage the strengths of both MSLMs and LVLMs, we propose the **E**xplainable **A**daptive **R**ationale-**A**ugmented **M**ultimodal (EARAM) fake news detection framework. This novel framework combines the task-specific integration and classification capabilities of MSLMs with the extensive commonsense knowledge and logical reasoning abilities of LVLMs. Specifically, we first utilize Retrieval-Augmented LVLMs to analyze news content from two perspectives: common sense and image-text complementarity logic. Then, MSLMs extract useful rationales from the generated analyses based on the original news text-image features and perform the judgment task. Finally, the reasons and judgment results from the MSLMs are fed back into the LVLMs to generate reliable explanations. This process fully leverages the respective advantages of MSLMs and LVLMs, thereby enabling the model to achieve higher accuracy while also having strong generalization and explainability.

Our main contributions are as follows:

- Comprehensive Evaluation: We evaluated the capabilities of common LVLMs in fake news detection, analyzed their limitations in direct decision-making, and proposed the idea of using LVLMs to generate multi-perspective analyses of news content.
- Novel and Practical Framework: We introduce the EARAM framework, which leverages the strengths of both MSLMs and LVLMs, merging the task-specific integration and classification prowess of the former with the breadth of knowledge and depth of language understanding of the latter. To the best of our knowledge, this is the first exploration of its kind in the multimodal field.
- Extensive Experimental Validation: We conduct comprehensive experiments on established benchmark datasets, rigorously demonstrating the superior performance of EARAM in terms of accuracy, generalization, and explainability.

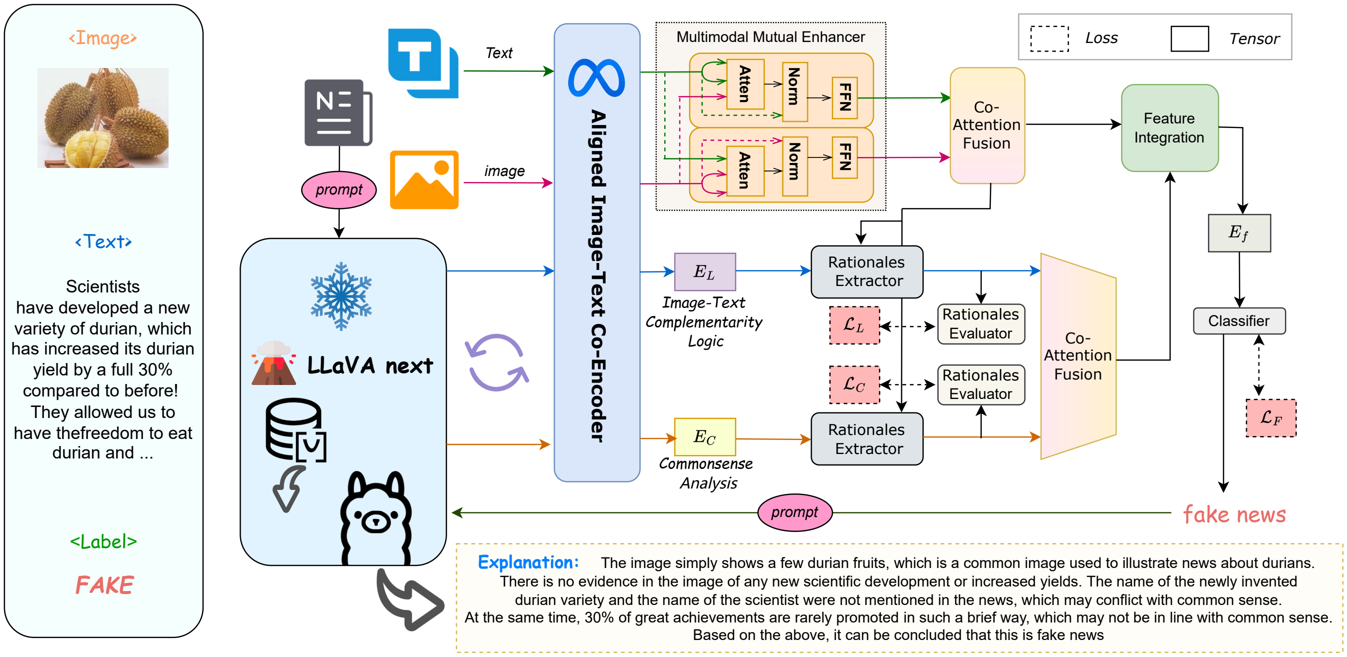
2 Related Work

2.1 Multimodal Fake News Detection

Previous research on multimodal fake news detection has largely been based on supervised task-specific Multimodal Small Language Models (MSLMs), primarily following two approaches [12].

The first is a content-based method [31], which solely utilizes the image and text information of the news [3, 32, 50]. Similar methods either train models from scratch or leverage pre-trained models, such as ViT [7] or RoBERTa [23], to represent text and image as feature vectors for subsequent classification tasks. However, they perform well only on the test set, which has a similar distribution to the training data [28], resulting in poor generalization ability [10]. They also lack readable explanations [43]. The second is a context-based method [12], which is typically based on graph neural networks (GNNs) [13, 26, 48] and additional information such as the news propagation tree [24], comments [54], and the identity of the publisher to assist in the judgment [35]. However, not all news articles can immediately access this information, resulting in high latency, which makes it difficult to detect fake news in real time [37].

We believe that these approaches based on MSLMs, which train additional task-specific classification layers on certain datasets,



Scientists have developed a new variety of durian, which has increased its durian yield by a full 30% compared to before! They allowed us to have thefreedom to eat durian and ...

Explanation: The image simply shows a few durian fruits, which is a common image used to illustrate news about durians. There is no evidence in the image of any new scientific development or increased yields. The name of the newly invented durian variety and the name of the scientist were not mentioned in the news, which may conflict with common sense. At the same time, 30% of great achievements are rarely promoted in such a brief way, which may not be in line with common sense. Based on the above, it can be concluded that this is fake news

Figure 2: Illustration of EARAM framework. For news articles composed of image-text pairs, the EARAM framework employs a three-step processing method: (a) Multi-perspective analyses of news content are generated by retrieval-augmented LVLMs, and then the news images, text, and analysis content are processed by image-text co-encoder. (b) The content between different modalities is allowed to fully interact and selectively extract useful rationales from the analysis. We then aggregate them and make the final judgment. (c) Finally, the rationales and prediction results will be input into the LVLMs to make a reasonable explanation.

oversimplify the problem into an end-to-end classification model [37]. These methods struggle to understand the content and logic of the news deeply, merely identifying surface signals conveyed through text and images [21].

2.2 Large Vision-Language Models

Recently, the rapid advancements in LLMs and LVLMs have greatly improved their ability to understand multimodal data and handle complex reasoning tasks [17–20, 44, 55, 56]. Some studies have begun exploring the use of LLMs and LVLMs for fake news detection [38]. Cheung and Lam [5] employ LoRA tuning to train a detector based on LLaMA, combining it with external retrieved knowledge. However, this method primarily focuses on detecting fake news content and lacks the ability to provide easily understandable explanations to the public. Hu et al. [9] find that LLMs perform poorly in truthfulness judgments, still exhibiting a significant gap compared to fine-tuned small language models like BERT. Nevertheless, they acknowledge the broad knowledge possessed by LLMs as essential. To address this, they adopt a knowledge distillation framework to train small language models that adaptively derive insights from the reasoning generated by LLMs. Their discussion, however, is limited to text. Wan et al. [42] decompose the fake news detection task into smaller sub-tasks, utilizing LVLMs to complete these sub-tasks and ultimately integrating the results. Their approach shows certain improvements compared to using CoT and self-consistency methods [47] on LVLMs alone, but there is still a gap in accuracy compared to MSLM-based methods.

In contrast to the above approaches, our model achieves explainability and high generalization while maintaining satisfactory accuracy, fully leveraging the strengths of both LVLMs and MSLMs.

3 Explainable Adaptive Rationale-Augmented Multimodal Framework

In this section, we present the novel and practical **E**xplainable **A**daptive **R**ationale-**A**ugmented **M**ultimodal (EARAM) Framework. The framework is primarily divided into three modules: **1) Generation and Representation:** The purpose of this module is to complete preliminary work. It employs retrieval-augmented Large Vision-Language Models to generate multi-perspective analyses of news content. Subsequently, both the analyses and the original news content are represented as feature tensors for further operations. **2) Interaction and Extraction:** This module aims to enable full interaction between the semantics of the text and images in the news, as well as adaptively extract useful rationales from analyses based on the original news. **3) Decision and Explanation:** In this stage, we use a feature integration module to conduct the aggregation and prediction task. The prediction results and extracted rationales are then input back into the LVLMs in the form of prompts to generate the final explanation.

3.1 Generation and Representation

3.1.1 ***Retrieval Augmentation.*** We aim to reduce hallucinations and enable LVLMs to access the latest corpus and data to address the timeliness of news. Since real-time training of models to learn the latest events is not feasible, we introduce the Retrieval-Augmented Generation (RAG) method [25]. We use official BBC news from 2017 to 2024 as the model's knowledge base[1], which has sufficient authority and can be considered as reliable source. When calculating similarity, we use the ensemble similarity calculation method, combining cosine similarity and BM25 similarity to create a more comprehensive and accurate similarity measure [15].

$$\begin{aligned} Similarity(N, T^i) &= \gamma_1 \cdot \frac{N_v \cdot T_v^i}{\|N_v\| \|T_v^i\|} \\ &+ \gamma_2 \cdot \sum_{j=1}^{n} \mathrm{IDF}(q_j) \cdot \frac{f(q_j, T^i) \cdot (k_1 + 1)}{f(q_j, T^i) + k_1 \cdot \left(1 - b + b \cdot \frac{|T^i|}{avg\bar{l}_d}\right)}, \end{aligned} \tag{1}$$

for each word q_j in query N, $f(q_j, T^i)$ represents its term frequency in the document. T_v^i and N_v are embedding vectors for the text chunk and news(denoted by the subscript v to indicate vectorized), respectively. $avg\bar{l}_d$ is the average document length, while k_1, b, γ_1, and γ_2 are adjustable parameters. $\mathrm{IDF}(q_i)$ denotes inverse document frequency [27]. The system ranks text chunks by similarity scores and selects the top three most similar knowledge blocks as retrieval results. To handle outdated or conflicting content, we store timestamps in the database and integrate them into the prompt. LVLMs will prioritize the most recent content, addressing potential inconsistencies. In actual deployment, we can use a real-time updated database for RAG, with data sourced from authoritative official media. This would enable the RAG module to achieve better performance than in the experimental environment.

3.1.2 ***Analyses Generation.*** Combining common-sense analysis and image-text complementarity logic is highly effective in assessing news authenticity [9]. The common-sense analysis evaluates news against established cognitive patterns in social, cultural, and scientific contexts, offering a macro perspective for initial screening. Meanwhile, image-text complementarity logic focuses on internal consistency between text and visuals, identifying micro-level inconsistencies, while also examining how the images and text can corroborate each other to strengthen the overall credibility of the news. Together, these perspectives provide a comprehensive, efficient evaluation of news authenticity from both macro and micro levels. Based on these two perspectives, we construct the input for the LVLMs through specific prompt engineering:

$$input = \Psi(x, v, P_i, Q), \tag{2}$$

where x represents the news text, v represents the news image, P_i represents the prompt instructing LVLMs to generate analysis from specific perspectives, Q represents the similar text blocks retrieved from the vector database, and Ψ represents prompt engineering techniques. After being input into the LVLMs, the analyses r_L and r_C are generated from the perspectives of image-text complementarity logic and common sense, respectively. The joint embedding space of MetaCLIP captures rich semantic information, which facilitates subsequent cross-modal tasks. We bypass the pooling operations and image-text similarity calculations of MetaCLIP. Instead, we use CLIP's text encoder, ViT image encoder, and the final projection layers to obtain the tensor representations of x, v, r_L, and r_C as $\mathbf{E_X}, \mathbf{E_V}, \mathbf{E_L}$ and $\mathbf{E_C}$, all with the same hidden layer dimensions. The discussion and details regarding the prompts are presented in the appendix §B.

3.2 Interaction and Extraction

3.2.1 ***Multimodal Mutual Enhancer.*** Fake and true news often differ semantically in both images and text. For example, true news tends to be objective and neutral, while fake news often uses strong emotional stimuli, exaggerated language, and specific patterns to increase virality. Although MetaCLIP excels at general image-text understanding, it may miss these subtle differences. To address this, we introduced the Multimodal Mutual Enhancer, which uses multi-head cross attention to better capture the specific semantic features that distinguish fake news from true news, improving detection accuracy and reliability.

Let $\mathbf{E_X}$ and $\mathbf{E_V}$ be the input tensors for news content and image, H be the number of heads in multi-head cross-attention, d be the model dimension, and $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the query, key, and value in the attention mechanism. ${\mathbf{Q}^{(1)}}_h, {\mathbf{K}^{(2)}}_h, {\mathbf{V}^{(2)}}_h$ are obtained by calculating $\mathbf{E_X}$ and $\mathbf{E_V}$ with parameter matrices. Then we calculate the enhanced results for each modality:

$$\begin{aligned} \mathbf{Z}^{(1)} &= \mathrm{Attention}(\mathbf{E_X}, \mathbf{E_V}, \mathbf{E_V}) \\ &= \left(\Big\|_{h=1}^{H} \mathrm{softmax}\left(\frac{{\mathbf{Q}^{(1)}}_h {\mathbf{K}^{(2)}}_h^{\top}}{\sqrt{d/H}} \right) {\mathbf{V}^{(2)}}_h \right) {\mathbf{W}^{(1)}}_O \Bigg), \end{aligned} \tag{3}$$

where "||" denotes the concatenation operation. ${\mathbf{W}^{(1)}}_O, {\mathbf{W}^{(2)}}_O \in \mathbb{R}^{d \times d}$ are the final linear transformation layers in the model. Similarly, $\mathbf{Z}^{(2)}$ can be obtained. Then, we perform residual connection through LayerNorm:

$$\begin{aligned} \mathbf{E_{X \leftarrow V}} &= \mathrm{LayerNorm}(\mathbf{E_X} + \mathbf{Z}^{(2)}), \\ \mathbf{E_{V \leftarrow X}} &= \mathrm{LayerNorm}(\mathbf{E_V} + \mathbf{Z}^{(1)}), \end{aligned} \tag{4}$$

$\mathbf{E_{X \leftarrow V}}$ and $\mathbf{E_{V \leftarrow X}}$ represent the enhanced news text and image representation tensors respectively.

3.2.2 ***Co-Attention Fusion.*** Co-Attention is a variant of standard multi-head self-attention that can capture global dependencies across all positions in a sequence, enhancing model performance by mutually focusing on important features of each other. First, through adaptive average pooling, the text features are average-pooled to the same dimension as the image features, facilitating subsequent fusion operations. At the same time, important features in both image and text are obtained through attention.

$$\begin{aligned} \mathbf{E'_X} &= \mathrm{Adaptive\_pool}(\mathbf{E_{X \leftarrow V}}), \\ \mathbf{f_1} &= \mathrm{Attention}(\mathbf{E'_X}, \mathbf{E_{V \leftarrow X}}, \mathbf{E_{V \leftarrow X}}), \end{aligned} \tag{5}$$

where $\mathbf{E'_x}$ and $\mathbf{E_{V \leftarrow X}}$ have the same dimensions. Similarly, $\mathbf{f_2}$ can be obtained using the same formula. Then, Co-Attention fusion is performed to combine important semantic features from text and

[1] https://huggingface.co/datasets/RealTimeData/bbc_news_alltime

image according to learnable weights:

$$\begin{aligned} \mathbf{R_t} &= w_1 \cdot \mathbf{f}_1 + w_2 \cdot \mathbf{f}_2, \\ \mathbf{R_p} &= \text{Avg\_pool}(\mathbf{R_t}), \end{aligned} \tag{6}$$

both $\mathbf{R_t}$ and $\mathbf{R_p}$ are representations of the overall semantics of the multimodal news. w is a learnable weight. However, $\mathbf{R_t}$ has the same dimensions as $\mathbf{E_{V \leftarrow X}}$ and will be used in the Rationale Extractor module. $\mathbf{R_p}$ is the result of average pooling $\mathbf{R_t}$ along the last dimension, reducing its dimensionality, and will be used in the final aggregation and judgment module.

3.2.3 *Rationales Extractor and Evaluator*.

The analyses directly generated by LVLMs for news are not always helpful in correctly determining the authenticity of the news. Since not all content is equally important or relevant [37], we adaptively extract rationales from the analysis based on the semantic features of the original news content that help us make correct judgments:

$$\begin{aligned} \mathbf{E'_L} &= \text{Attention}(\mathbf{R_t}, \mathbf{E_L}, \mathbf{E_L}), \\ \mathbf{E'_C} &= \text{Attention}(\mathbf{R_t}, \mathbf{E_C}, \mathbf{E_C}), \\ \mathbf{R'_A} &= \text{Fusion}(\mathbf{E'_C}, \mathbf{E'_L}), \end{aligned} \tag{7}$$

where $\mathbf{R'_A}$ is the aggregation of multi-perspective analyses, using the same aggregation method as Co-Attention fusion. To more accurately evaluate the usefulness of the information extracted from the analyses, we have manually designed a Rationales Evaluator module consisting of a feedforward neural network that predicts news classification solely based on the analyses. The network consists of 3 hidden layers with corresponding ReLU activation functions.

$$\begin{aligned} &\hat{\mathbf{y}}_L = \text{FFN}(\mathbf{E_L}) \ , \ \hat{\mathbf{y}}_C = \text{FFN}(\mathbf{E_C}), \\ &\mathcal{L}_L = -\mathbf{y} \log(\hat{\mathbf{y}}_L) - (1 - \mathbf{y}) \log(1 - \hat{\mathbf{y}}_L), \\ &\mathcal{L}_C = -\mathbf{y} \log(\hat{\mathbf{y}}_C) - (1 - \mathbf{y}) \log(1 - \hat{\mathbf{y}}_C), \end{aligned} \tag{8}$$

where $\mathbf{y}$ is the ground truth, and $\mathcal{L}_C$ and $\mathcal{L}_L$ are the cross-entropy loss functions between the predicted values and the ground truth.

It should be noted that there is useless information in the analysis from LVLMs, and there may even be conflicts between analyses from different perspectives. To address this, we implement a cross-attention mechanism to focus on content in the generated analysis that is more relevant to the original news features and extract rationales from it. To mitigate potential conflicts between these rationales, we separately set loss functions $\mathcal{L}_C$ and $\mathcal{L}_L$, which are used to evaluate the contribution of the extracted rationales to the final correct prediction, training the Rationales Extractor module to filter out noise.

The Fusion and Integration modules in the model are used to weigh the contributions of different rationales. Metaphorically speaking, these modules can be viewed as a 'judge', whose purpose is to handle conflicting arguments and determine the 'winner' among them. This approach of using small models as judges to resolve conflicting viewpoints has also demonstrated effectiveness in previous research [21].

3.3 Decision and Explanation

Based on the previous outputs, we aggregate the semantic features of the news content and the filtered multi-perspective rationales together for the final judgment. λ is a learnable parameter.

$$\mathbf{E_F} = \lambda \cdot \mathbf{R_p} + (1 - \lambda) \cdot \mathbf{R_A}, \tag{9}$$

then, through a classifier composed of a feedforward neural network with multiple hidden layers and activation functions, we obtain the classification result $\hat{\mathbf{y}}_F$. Then we compare the prediction with the ground truth using the cross-entropy loss function:

$$\begin{aligned} &\hat{\mathbf{y}}_F = \text{FFN}(\mathbf{E_F}), \\ &\mathcal{L}_F = -\mathbf{y} \log(\hat{\mathbf{y}}_F) - (1 - \mathbf{y}) \log(1 - \hat{\mathbf{y}}_F). \end{aligned} \tag{10}$$

After completing news judgment tasks, the LVLMs are used again to obtain the final reliable explanation. We effectively utilize the prediction results of the MSLMs and the analyses previously generated by the LVLMs to enable the LVLMs to produce more accurate and reliable analyses.

$$\begin{aligned} r'_i &= \begin{cases} r_i, & \text{if} \quad \hat{y}_F = y_i \\ \text{Re-gen}(r_i, \hat{y}_F), & \text{else} \end{cases} \\ input &= \Psi(x, v, \hat{y}_F, r'_L, r'_C, P), \end{aligned} \tag{11}$$

where $Re - gen$ indicates regenerating specific analyses of the news from that perspective.

The final loss function is a weighted sum of the previous loss functions:

$$\mathcal{L} = \mathcal{L}_F + \alpha \cdot \mathcal{L}_L + \beta \cdot \mathcal{L}_C + \gamma \cdot \|\boldsymbol{w}\|^2, \tag{12}$$

where α and β are hyperparameters used to balance the two loss functions. γ is the hyperparameter for weight decay, used to control the strength of regularization.

4 Experiments

The experiments are divided into five parts. We first evaluate the capabilities of LVLMs, exploring whether LVLMs can be directly used to judge the truthfulness of news. We then focus on evaluating the EARAM model, comparing its performance with existing models on benchmark datasets to assess EARAM's ability to detect fake news. After that, we conduct ablation studies on the model, examining the importance of each module within the model. Subsequently, we focus on the quality of the explanations generated by the model. Finally, we examine the generalization capability of the model, exploring the advantages of EARAM in terms of generalization and transferability compared to existing methods.

Additionally, we conduct a case study, which can be found in Appendix §C, to provide a more comprehensive understanding of the method and highlight its benefits.

4.1 Experimental Settings

Dataset. Our experiments involve three real-world multimodal datasets: Weibo [2], Pheme [16], and MR2-en [14]. Our main experiments are conducted on the most widely used datasets, the Chinese dataset Weibo and the English dataset Pheme. In the generalization experiments and explanation generation experiments, to ensure fairness by avoiding the impact of language differences, we introduced the MR2-en dataset. We removed data labeled as '*Unverified*' from the original MR2-en dataset. For Pheme and Weibo datasets, we preprocessed and divided the data following the methods used in MCAN [50] and MMCAN [59], respectively.

Table 1: Evaluation of LVLMs capabilities. The accuracy is reported as the metrics.

Model	Usage	Pheme	MR2
InternVL2-8B	zero-shot	0.4113	0.5298
LLaVA-Next -mistral-7b	zero-shot	0.4823	0.5455
	Few-Shot	0.5028	0.5486
	Zero-Shot CoT	0.5112	0.6144
	Few-Shot CoT	0.6023	0.6771
	Self-Consistency	0.6536	0.7048
	Multi-Agent Debate	0.6791	0.6942
MCAN	-	0.8650	0.8973

Baseline. To validate the effectiveness of our model, we selected several state-of-the-art unimodal and multimodal methods for comparison. The unimodal methods include: 1) **BERT** [6]; 2) **DeiT** [40]; The multimodal methods include: 3) **CAFE** [3]; 4) **HMCAN** [32]; 5) **MCAN** [50]; 6) **MMCAN** [59]; 7) **COOLANT** [45]; The evaluation metrics include accuracy, precision, recall, and F1 score.

The data statistics, baseline descriptions, and model implementation are detailed in Appendix §A.[2]

4.2 Evaluation of LVLMs Capabilities

From the results in Table 1, we can observe that directly using LVLMs for fake news detection yields poor performance, even when employing chain-of-thought (CoT) or few-shot methods [49]. These approaches underperform compared to classic supervised task-specific multimodal small language models trained from scratch, such as MCAN. Additionally, we found that CoT significantly outperforms few-shot. This may be because the labels in the examples only contain classification results of 0 or 1, which do not provide much useful information to LVLMs when using few-shot prompts. In contrast, the few-shot CoT examples we carefully selected include three instances with step-by-step reasoning and inductive processes, which evidently provide more information to the model, enabling it to learn how to reason towards an answer incrementally.

Self-Consistency [47] enhances the model's stability and accuracy in uncertain tasks by performing multiple model inferences and result voting. Multi-Agent Debate [8] is an adversarial debate-based training method that improves the model's reasoning ability and decision quality through discussions between multiple agents. These are both recognized as cutting-edge methods, yet they still fail to enable LVLMs to match the judgment capability of MSLMs.

In conclusion, directly using LVLMs for fake news detection is indeed unsatisfactory. This drives us to explore alternative methods to more effectively leverage the vast knowledge and understanding capabilities that LVLMs possess.

4.3 Performance of EARAM Model

In Table 2, we compare the EARAM model with existing unimodal and multimodal fake news detection models in terms of detection accuracy. We can observe the following key points: 1). Unimodal methods perform significantly worse than multimodal methods overall. This is primarily because unimodal methods are limited to features from a single modality (such as text or image), while multimodal models capture a broader range of semantic information by integrating both visual and textual features. 2). On the Weibo and Pheme datasets, the EARAM model outperforms all baseline models in terms of accuracy. On the Pheme dataset, the accuracy and F1 score for fake news detection improved by 1.5% and 2.1%, respectively, compared to the previously best-performing model. This indicates that leveraging LVLMs' capabilities can further enhance the accuracy of fake news detection. 3). The overall accuracy on the Weibo dataset is lower than on Pheme, especially for unimodal methods relying only on images. This suggests that text plays a more significant role in the Weibo dataset.

From these results, it is evident that the EARAM model consistently outperforms baseline models on both the Weibo and Pheme datasets, demonstrating that utilizing LVLMs' abilities can further improve the accuracy of fake news detection.

4.4 Ablation Study

In the ablation study, to further investigate the effectiveness of each module of the EARAM model, we perform a quantitative analysis by removing each module and comparing it with the following variants: *-w/o Image*: This variant only uses the news text as input and removes image content. *-w/o Text*: This variant only uses the news image as input, removing the text content. Both of these variants only use unimodal content, and the modality interaction fusion step in the Multimodal Mutual Enhanced (MME) module is disabled. *-w/o LVLMs*: This variant discards the output from LVLMs during the final feature aggregation process. *-w/o EE*: This variant discards the Evidence Extraction (EE) module and directly concatenates the LVLMs' analysis with the feature representation of the news content for classification. *-w/o RAG*: This variant does not use the retrieval-augmented generation (RAG) mechanism in the LVLMs analysis phase, meaning it does not perform retrieval from our constructed database. *-w/o CS*: In this variant, the analysis phase does not consider reasoning from a common sense perspective. *-w/o CL*: In this variant, the analysis phase does not involve reasoning from the image-text complementary logic perspective. *-w/o MME*: This variant removes the Multimodal Mutual Enhanced (MME) module and directly concatenates the E_L and E_C tensors for classification.

From the results in table 3, it is evident that the *w/o Image* and *w/o Text* variants, which use only unimodal content, perform the worst. This demonstrates the necessity of integrating multimodal features in EARAM to make a comprehensive judgment. The *w/o MME* variant performs poorly because it lacks the original news features, preventing the evidence extraction module from effectively extracting useful information from the analysis. This also leads to a loss of information during the final aggregation. The *w/o LVLMs* variant shows a performance gap of about 6% to 9% compared to EARAM, indicating that LVLMs significantly contribute to assisting EARAM in making the correct final judgment. The performance of the *w/o RAG* variant is 4% lower compared to EARAM in Pheme, indicating that the knowledge retrieved from BBC News helps EARAM more effectively identify fake news. The accuracy drop in both the *w/o CS* and *w/o CL* variants is around 6% to 8%, and it is worth noting that the performance of the *w/o CS* variant is better than that of *w/o CL*. This indicates that, in the final judgment,

[2] The source code is available at https://github.com/qingpingwan/EARAM

Table 2: Performance comparison between EARAM model and other models. The best results are in bold.

Datasets	Models	Accuracy	Fake News			Real News		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	BERT	0.722	0.717	0.706	0.711	0.725	0.769	0.746
	DeiT	0.656	0.604	0.668	0.634	0.697	0.680	0.688
	CAFE	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	HMCAN	0.885	0.920	0.845	0.881	0.856	0.926	0.890
	MCAN	0.899	0.913	0.889	0.901	0.884	0.909	0.897
	MMCAN	0.911	0.913	0.910	0.912	0.909	0.912	0.911
	COOLANT	0.923	0.927	**0.923**	**0.925**	0.919	0.922	0.920
	EARAM	**0.9310**	**0.9421**	0.9062	0.9238	**0.9221**	**0.9523**	**0.9370**
Pheme	BERT	0.743	0.735	0.726	0.730	0.778	0.759	0.768
	DeiT	0.738	0.721	0.742	0.731	0.754	0.737	0.745
	CAFE	0.861	0.812	0.645	0.719	0.875	0.943	0.908
	HMCAN	0.881	0.830	0.838	0.834	0.910	0.905	0.907
	MCAN	0.865	0.790	0.680	0.731	0.887	0.933	0.910
	MMCAN	0.903	0.855	0.777	0.814	0.918	**0.950**	0.934
	COOLANT	0.926	**0.886**	0.853	0.869	0.942	0.945	0.943
	EARAM	**0.9406**	0.8824	**0.8982**	**0.8902**	**0.9652**	0.9480	**0.9565**

Table 3: Ablation studies by removing modules from our proposed framework.

Models	Pheme		Weibo	
	Acc	F1	Acc	F1
EARAM	0.940	0.890	0.931	0.923
-w/o Image	0.692	0.667	0.725	0.734
-w/o Text	0.744	0.725	0.759	0.751
-w/o LVLMs	0.853	0.832	0.841	0.834
-w/o EE	0.842	0.818	0.820	0.807
-w/o RAG	0.904	0.852	0.913	0.894
-w/o CS	0.896	0.879	0.881	0.875
-w/o CL	0.865	0.854	0.870	0.871
-w/o MME	0.808	0.795	0.786	0.784

reasoning about the inherent complementary logic between images and text, and exploring how they corroborate each other, is more beneficial for making the correct prediction. This also suggests that we could try to find more effective prompts to guide the model in analyzing news content. The *w/o EE* variant performs 10% to 11% worse than EARAM, which highlights that the Rationales Extractor (RE) module effectively extracts relevant and useful rationales from the raw analysis generated by LVLMs. It also shows that the raw analysis from LVLMs contains a certain amount of noise, which can impact the final judgment, emphasizing the need to filter out irrelevant features.

4.5 Evaluations on Explanation

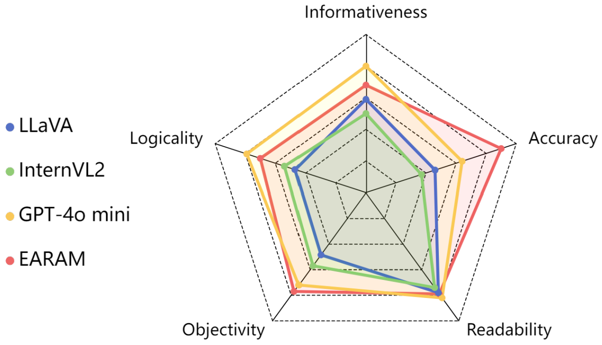
To evaluate the explainability of the models, we opted for a human subjective study, as automatic evaluation methods fall short when assessing the quality of generated explanations due to the diverse nature of textual expressions and the lack of a gold standard for explanations in the fake news detection task. Even when we attempted to use GPT-4o for evaluation, we found that the scores provided by the GPT-4o were highly influenced by prompt design and input formatting [46, 60]. For instance, simply switching the order of explanations generated by different models in the input would affect GPT-4o's final scoring. Additionally, GPT-4o often tended to give higher ratings to explanations generated by GPT-4o mini, even when the explanations were incorrect.

In light of these challenges, we implement a rigorous human subjective evaluation methodology. This approach, known as the **ILORA** Explanation Quality Evaluation Method [62], employs a 5-point Likert scale to assess the overall quality of generated explanations across five critical dimensions:

- **Informativeness (I)**: Evaluating whether the explanation provides new information, such as background or additional context.
- **Logicality (L)**: Assessing whether the explanation follows a reasonable thought process and whether there is a strong causal relationship between the explanation and the result.
- **Objectivity (O)**: Evaluating whether the explanation is objective and free from excessive subjective emotion.
- **Readability (R)**: Assessing whether the explanation follows proper grammatical and structural rules, with coherent sentences that are easy to understand.
- **Accuracy (A)**: Evaluating whether the generated explanation aligns with the actual truth label and whether the explanation accurately reflects the result.

Each criterion is rated on a 5-point scale, where 1 indicates the lowest quality and 5 indicates the highest quality, allowing for a nuanced assessment. We select 10 volunteers to randomly rate 60 explanations drawn from the MR2-en dataset. This sample size is chosen to balance comprehensive coverage with practical feasibility. For comparison, we use three state-of-the-art models: LLaVA-Next [17], InternVL2 [4], and GPT-4o mini [30] to generate explanations for fake news. This selection of models represents a spectrum of current capabilities in multimodal and language understanding, providing a comprehensive benchmark for our evaluation.

Figure 3: Evaluation Results of Explanations Generated by Four Models on the MR2 Dataset Across Five Dimensions.



To maintain fairness, we explicitly specify in the prompts that the generated responses should be around 50 words, aiming to make the average response length similar across the four models. Meanwhile, LLaVA-Next, InternVL2, and GPT4o-mini use the Chain of Thought (CoT) technique to enhance the output quality, while the EARAM model does not employ the RAG technique to improve its output. By assessing these explanations across five dimensions, we systematically evaluate the quality of the generated content and provide valuable feedback for improving model performance.

The results in figure 3 showed the following: 1) Explanations from all models score relatively high in the readability dimension because LVLMs are capable of generating fluent sentences. 2) The explanations generated by EARAM outperform those from LLava and InternVL2 across all dimensions, particularly in accuracy, indicating that our approach fully leveraged LVLMs' ability to detect fake news and provide explanations. 3) EARAM performs slightly worse than GPT-4o mini in terms of logicality and informativeness, likely due to GPT-4o mini's significantly larger parameter size. However, EARAM is stronger in accuracy, demonstrating its ability to achieve higher performance with fewer parameters, particularly in generating accurate explanations.

Overall, although there is still room for further improvement in the evaluation of explanations for fake news detection, the results suggest that leveraging the advanced capabilities of LVLMs to design a general framework for both fake news detection and explanation is feasible and effective. Additionally, case study on the generated explanations is also included in the appendix §C.

4.6 Generalization Research

To evaluate the model's generalization ability across datasets, we designed a set of experiments comparing our proposed EARAM model with the two best-performing supervised task-specific MSLMs methods. In these experiments, we primarily explore how the models perform on new datasets to assess their generalization and transferability. To ensure the fairness of the experiments and eliminate the impact of language differences, we do not use the Chinese Weibo dataset but introduce a new English dataset, MR2 [14]. This approach is more conducive to observing the models' adaptability to new classification tasks. In the transfer learning experiments, we froze all model parameters except for the last MLP classification head. With this setup, we aim to focus on evaluating the model's transferability, ensuring that it relies on learned feature representations for transfer, rather than retraining large-scale parameters.

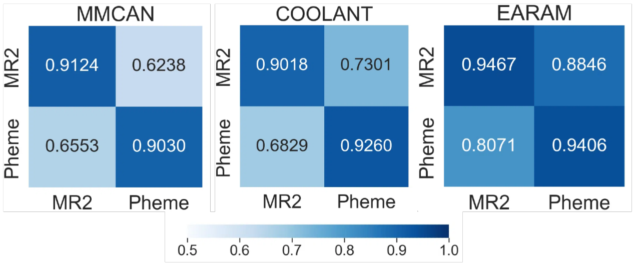


Figure 4: Generalization Experiments on the Pheme and MR2 Datasets. The vertical axis represents the training set, and the horizontal axis represents the test set. The accuracy is reported as the metrics.

Figure 4 shows the experimental results. By comparison, it can be observed that when the parameters of all models are frozen, the test results of supervised task-specific MSLMs on the new dataset drop sharply compared to EARAM. The accuracy of MMCAN on the Pheme test set drops by nearly 30%. This indicates that the EARAM model significantly outperforms MMCAN and COOLANT in terms of generalization and transferability. The EARAM model is able to more effectively transfer learned data representations to new datasets, demonstrating its superior feature representation capabilities and enhanced generalization performance. Additionally, the results suggest that models trained on the MR2 dataset exhibit better generalization and transferability, likely due to the higher quality of the MR2 dataset, which better captures the underlying features of fake news.

The experimental results not only prove the outstanding performance of the EARAM model on the original tasks but also demonstrate its adaptability in cross-dataset transfer scenarios, providing important insights for real-world applications in multi-task learning and cross-domain task processing.

5 Conclusion

In this paper, we propose a novel and practical EARAM model for fake news detection. EARAM integrates the flexible learning capabilities of supervised Multimodal Small Language Models (MSLMs) for specific tasks with the vast common sense and logical reasoning abilities of Large Vision-Language Models (LVLMs). By combining MSLMs and LVLMs, our method not only enhances the model's multimodal fake news detection capability but also improves the model's generalization and explainability. Extensive experiments on three public benchmark datasets also demonstrate the effectiveness of our model.

Acknowledgment

This work was supported by the National Nature Science Foundation of China (No. 62192781, No. 62272374), the Natural Science Foundation of Shaanxi Province (2024JC-JCQN-62), the National Nature Science Foundation of China (No. 62202367, No. 62250009, No. 62137002), the Key Research and Development Project in Shaanxi Province No. 2022GXLH-01-03, Project of China Knowledge Center for Engineering Science and Technology, and Project of Chinese academy of engineering "The Online and Offline Mixed Educational Service System for 'The Belt and Road' Training in MOOC China". We would like to express our gratitude for the support of K. C. Wong Education Foundation. We also appreciate the reviewers and chairs for their constructive feedback.

Lastly, we would like to thank all LUD lab members for fostering a collaborative research environment.

References

[1] Fabrício Benevenuto and Philipe Melo. 2024. Misinformation Campaigns through WhatsApp and Telegram in Presidential Elections in Brazil. *Commun. ACM* 67, 8 (Aug. 2024), 72–77. doi:10.1145/3653325

[2] Qi Cao, Huawei Shen, Keting Cen, Wentao Ouyang, and Xueqi Cheng. 2017. DeepHawkes: Bridging the Gap between Prediction and Understanding of Information Cascades. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore, Singapore) *(CIKM '17)*. Association for Computing Machinery, New York, NY, USA, 1149–1158. doi:10.1145/3132847.3132973

[3] Yixuan Chen et al. 2022. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In *The ACM Web Conference*. 2897–2905.

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *arXiv preprint arXiv:2312.14238* (2023).

[5] Tsun-Hin Cheung and Kin-Man Lam. 2023. FactLLaMA: Optimizing Instruction-Following Language Models with External Knowledge for Automated Fact-Checking. arXiv:2309.00240 [cs.CL] https://arxiv.org/abs/2309.00240

[6] 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, Volume 1 (Long and Short Papers)*. ACL, 4171–4186. doi:10.18653/v1/N19-1423

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929* (2021). https://arxiv.org/pdf/2010.11929

[8] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv:2305.14325 [cs.CL] https://arxiv.org/abs/2305.14325

[9] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 20 (2024), 22105–22113.

[10] Beizhe Hu, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Zhengjia Wang, and Zhiwei Jin. 2023. Learn over Past, Evolve for Future: Forecasting Temporal Trends for Fake News Detection. arXiv:2306.14728 [cs.CL] https://arxiv.org/abs/2306.14728

[11] Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu. 2022. Deep learning for fake news detection: A comprehensive survey. *AI Open* 3 (2022), 133–155. doi:10.1016/j.aiopen.2022.09.001

[12] Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu. 2022. Deep learning for fake news detection: A comprehensive survey. *AI Open* 3 (2022), 133–155. doi:10.1016/j.aiopen.2022.09.001

[13] Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to The Knowledge: Graph Neural Fake News Detection with External Knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 754–763.

[14] Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S. Yu. 2023. MR2: A Benchmark for Multimodal Retrieval-Augmented Rumor Detection in Social Media. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2023). https://api.semanticscholar.org/CorpusID:259950031

[15] Kush Juvekar and Anupam Purwar. 2024. COS-Mix: Cosine Similarity and Distance Fusion for Improved Information Retrieval. arXiv:2406.00638 [cs.IR] https://arxiv.org/abs/2406.00638

[16] Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task Learning for Rumour Verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, Emily M. Bender, Leon Derczynski, and Pierre Isabelle (Eds.). Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3402–3413. https://aclanthology.org/C18-1288

[17] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-NeXT: Stronger LLMs Supercharge Multimodal Capabilities in the Wild. https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).

[19] Zhaowei Li, Wei Wang, YiQing Cai, Xu Qi, Pengyu Wang, Dong Zhang, Hang Song, Botian Jiang, Zhida Huang, and Tao Wang. 2024. UnifiedMLLM: Enabling Unified Representation for Multi-modal Multi-tasks With Large Language Model. arXiv:2408.02503 [cs.CL] https://arxiv.org/abs/2408.02503

[20] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, YiQing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, Zhida Huang, and Tao Wang. 2024. GroundingGPT: Language Enhanced Multi-modal Grounding Model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 6657–6678. doi:10.18653/v1/2024.acl-long.360

[21] Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards Explainable Harmful Meme Detection through Multimodal Debate between Large Language Models. *arXiv preprint arXiv:2401.13298* (2024).

[22] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35. doi:10.1145/3560815

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019). https://arxiv.org/pdf/1907.11692

[24] Yang Liu and Yi-fang Brook Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. 354–361.

[25] Zhiwei Liu, Kailai Yang, Qianqian Xie, Christine de Kock, Sophia Ananiadou, and Eduard Hovy. 2024. AEmoLLM: Retrieval Augmented LLMs for Cross-Domain Misinformation Detection Using In-Context Learning based on Emotional Information. arXiv:2406.11093 [cs.CL] https://arxiv.org/abs/2406.11093

[26] 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*. 505–514.

[27] Xing Han Lù. 2024. BM25S: Orders of magnitude faster lexical search via eager sparse scoring. arXiv:2407.03618 [cs.IR] https://arxiv.org/abs/2407.03618

[28] Yida Mu, Xingyi Song, Kalina Bontcheva, and Nikolaos Aletras. 2024. Examining the Limitations of Computational Rumor Detection Models Trained on Static Datasets. arXiv:2309.11576 [cs.CL] https://arxiv.org/abs/2309.11576

[29] 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. *CoRR* abs/2405.16631 (2024). http://dblp.uni-trier.de/db/journals/corr/corr2405.html#abs-2405-16631

[30] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023). https://arxiv.org/pdf/2303.08774

[31] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving Fake News Detection by Using an Entity-enhanced Framework to Fuse Diverse Multimodal Clues. In *Proceedings of the 29th ACM International Conference on Multimedia*. ACM, 1212–1220. doi:10.1145/3474085.3481548

[32] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical Multi-modal Contextual Attention Network for Fake News Detection. In *The International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event*. 153–162.

[33] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a General-Purpose Natural Language Processing Task Solver? arXiv:2302.06476 [cs.CL] https://arxiv.org/abs/2302.06476

[34] Yoel Roth. 2022. The vast majority of content we take action on for misinformation is identified proactively. https://twitter.com/yoyoel/status/1483094057471524867. Accessed: 2023-08-13.

[35] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* 19 (2017), 22–36. doi:10.1145/3137597.3137600

[36] Márcio Silva and Fabrício Benevenuto. 2021. COVID-19 ads as political weapon. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing* (Virtual Event, Republic of Korea) *(SAC '21)*. Association for Computing Machinery, New York, NY, USA, 1705–1710. doi:10.1145/3412841.3442043

[37] Jinyan Su, Claire Cardie, and Preslav Nakov. 2024. Adapting Fake News Detection to the Era of Large Language Models. arXiv:2311.04917 [cs.CL] https://arxiv.org/abs/2311.04917

[38] Sahar Tahmasebi, Eric Müller-Budack, and Ralph Ewerth. 2024. Multimodal Misinformation Detection using Large Vision-Language Models. arXiv:2407.14321 [cs.CL] https://arxiv.org/abs/2407.14321

[39] Samia Tasnim, Md Mahbub Hossain, and Hoimonty Mazumder. 2020. Impact of rumors and misinformation on COVID-19 in social media. *Journal of Preventive Medicine and Public Health* 53, 3 (2020), 171–174.

[40] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. 2021. Training data-efficient image transformers distillation through attention. In *International Conference on Machine Learning*, Vol. 139. 10347–10357.

[41] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151. doi:10.1126/science.aap9559

[42] 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* (2024).

[43] 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*. Association for Computing Machinery, New York, NY, USA, 2452–2463. doi:10.1145/3589334.3645471

[44] Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, Yutong Zhang, Zihao Wu, Zhengliang Liu, Tianyang Zhong, Bao Ge, Tuo Zhang, Ning Qiang, Xintao Hu, Xi Jiang, Xin Zhang, Wei Zhang, Dinggang Shen, Tianming Liu, and Shu Zhang. 2024. A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges Across Different Tasks. arXiv:2408.01319 [cs.AI] https://arxiv.org/abs/2408.01319

[45] Longzheng Wang, Chuang Zhang, Hongbo Xu, Yongxiu Xu, Xiaohan Xu, and Siqi Wang. 2023. Cross-modal Contrastive Learning for Multimodal Fake News Detection. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. ACM. doi:10.1145/3581783.3613850

[46] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large Language Models are not Fair Evaluators. arXiv:2305.17926 [cs.CL] https://arxiv.org/abs/2305.17926

[47] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171 [cs.CL] https://arxiv.org/abs/2203.11171

[48] Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. Fake News Detection via Knowledge-driven Multimodal Graph Convolutional Networks. In *Proceedings of the International Conference on Multimedia Retrieval*. 540–547.

[49] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, Vol. 35. Curran Associates, Inc., 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[50] 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*. 2560–2569.

[51] Junhao Xu, Longdi Xian, Zening Liu, Mingliang Chen, Qiuyang Yin, and Fenghua Song. 2024. The Future of Combating Rumors? Retrieval, Discrimination, and Generation. arXiv:2403.20204 [cs.AI] https://arxiv.org/abs/2403.20204

[52] Rongwu Xu, Brian S Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2023. The Earth is Flat because...: Investigating LLMs' Belief towards Misinformation via Persuasive Conversation. *arXiv preprint arXiv:2312.09085* (2023).

[53] Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R Fung, and Heng Ji. 2024. LEMMA: Towards LVLM-Enhanced Multimodal Misinformation Detection with External Knowledge Augmentation. *arXiv preprint arXiv:2402.11943* (2024).

[54] Yuta Yanagi, Ryohei Orihara, Yuichi Sei, Yasuyuki Tahara, and Akihiko Ohsuga. 2020. Fake News Detection with Generated Comments for News Articles. In *2020 IEEE 24th International Conference on Intelligent Engineering Systems*. 85–90. doi:10.1109/INES49302.2020.9147195

[55] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. MM-LLMs: Recent Advances in MultiModal Large Language Models. arXiv:2401.13601 [cs.CL] https://arxiv.org/abs/2401.13601

[56] Hangtao Zhang, Chenyu Zhu, Xianlong Wang, Ziqi Zhou, Changgan Yin, Minghui Li, Lulu Xue, Yichen Wang, Shengshan Hu, Aishan Liu, et al. 2025. BadRobot: Manipulating embodied LLMs in the physical world. In *Proceedings of the International Conference on Learning Representations (ICLR'25)*.

[57] Xuan Zhang and Wei Gao. 2023. Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method. arXiv:2310.00305 [cs.CL] https://arxiv.org/abs/2310.00305

[58] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223* (2023). https://arxiv.org/pdf/2303.18223

[59] Jiaqi Zheng et al. 2022. MFAN: Multi-modal Feature-enhanced Attention Networks for Rumor Detection. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. 2413–2419.

[60] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL] https://arxiv.org/abs/2306.05685

[61] Xiaofan Zheng, Minnan Luo, and Xinghao Wang. 2025. Unveiling Fake News with Adversarial Arguments Generated by Multimodal Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 7862–7869. https://aclanthology.org/2025.coling-main.526/

[62] Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics* 10, 5 (2021). doi:10.3390/electronics10050593

[63] Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022. Generalizing to the Future: Mitigating Entity Bias in Fake News Detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. doi:10.1145/3477495.3531816

A Implementation Details

A.1 Datasets

The statistics of the datasets are shown in Table 4, divided into English and Chinese datasets. The Chinese dataset, Weibo, contains fake news collected from May 2012 to January 2016, verified and compiled by Xinhua News Agency and Weibo. The English datasets are Pheme and MR2-en. The Pheme dataset is based on five breaking news events, with each event comprising a set of posts including a large number of texts and images. MR2-en is a three-class dataset crawled from Twitter and verified using the Google Fact Check Tools API.

A.2 Baseline

- **BERT** [6]: A pretrained language representation model with bidirectional contextual understanding. The BERT model relies solely on text for making judgments.
- **DeiT** [40]: An efficient visual Transformer model optimized through knowledge distillation techniques. The DeiT model relies solely on images to make judgments.
- **CAFE** [3]: Learns cross-modal ambiguity and adaptively aggregates multimodal and unimodal features for fake news detection.
- **HMCAN** [32]: Uses a hierarchical multimodal context attention network to fuse cross-modal and intra-modal relationships, jointly modeling multimodal context and hierarchical text semantics.
- **MCAN** [50]: Utilizes co-attention layers to learn and fuse cross-modal relationships between textual and visual features for fake news detection.

Table 4: The statistics of three datasets.

		Non-rumors	Rumors	Image
Chinese	Weibo	877	590	1467
English	Pheme	1928	590	2018
	MR2-en	2318	1418	3736

- **COOLANT** [45]: A cross-modal contrastive learning framework that learns cross-modal correlations with an attention-guided module for effective fake news detection.
- **MMCAN** [59]: Implements an image-text matching-aware co-attention mechanism to enhance multimodal fusion for better fake news detection.

A.3 Training Settings

Our LVLMs use LLaVA-v1.6-mistral-7b (also known as LLaVa-Next) and InternVL-8B. For RAG, we employ the Chroma vector database[3], with text embeddings generated using the BGE-base-v1.5 model[4]. The CLIP model uses MetaCLIP-b16[5] for English datasets and Chinese CLIP for Chinese datasets[6]. Due to LLaVA's weak Chinese generation capability, we set LLaVA to output in English even for Chinese datasets, then translate to Chinese using the DeepL API[7].

We use the AdamW optimizer with an initial learning rate of 2e-5, training for 40 epochs with early stopping to prevent overfitting. Our batch size is set to 32, with 8 attention heads (H = 8). The balancing hyperparameters α and β are both set to 0.4. The weight hyperparameters γ_1 and γ_2 are set to 0.7 and 0.3, respectively. To mitigate overfitting, we applied a dropout rate of 0.3 and a weight decay rate of 2e-3.

All code is implemented in PyTorch and runs on NVIDIA V100 32G GPUs. Evaluation metrics include accuracy, precision, recall, and F1 score.

B Prompt Details

Different prompts can guide LVLMs in analyzing and reflecting on news content from diversified perspectives. In our process of designing prompts for LVLMs, we experimented with a broad set of prompts from multiple angles. Through repeated trials and comparisons, we found that prompts framed from the perspectives of common sense reasoning and image-text complementary logic more effectively stimulated deeper thinking and higher-quality output from the model.

It is important to note that the possibilities for prompts are limitless, and we cannot claim that our designed prompts are the optimal solution [22]. However, through rigorous ablation experiments, we were able to demonstrate that our carefully crafted prompts significantly improved the quality of the model's output, helping it make more accurate judgments and inferences.

Nevertheless, we recognize that there is still vast room for exploration in this field. Future research may discover more effective prompt design strategies or methods like soft prompts that could be utilized to further enhance the quality of LVLM outputs.

B.1 Common Sense Analysis Prompt

"[INST] <image> news text: <text> information retrieved from authoritative news sources: <RAG retrieved content with timestamp> Please provide a comprehensive analysis of the given news text and image. From a common sense perspective, consider whether this news conflicts with widely accepted knowledge or established perceptions. These clues may be used with others to further predict the truth of the news, so your answer does not need to provide a definitive conclusion. Limit your response to 50 words. [/INST]"

B.2 Image-Text Complementarity Prompt

"[INST] <image> news text: <text> information retrieved from authoritative news sources: <RAG retrieved content with timestamp> Please provide a comprehensive analysis of the given news text and image. Examine the consistency between the text and the image, focusing on the internal coherence between the content and visual elements. These clues may be used with others to further predict the truth of the news, so your answer does not need to provide a definitive conclusion. Limit your response to 50 words. [/INST]"

B.3 Final Generated Explanation Prompt

"[INST] <image> news text: <text> preliminary analysis: <analysis 1> <analysis 2> We already know this news is <true/fake>. This is a preliminary analysis of the news from common-sense and image-text complementary perspectives before its truth was determined, which you may partially reference. Please provide reasons and explanations for why this news is <true/fake>, considering both the image and text. Limit your response to 50 words. [/INST]"

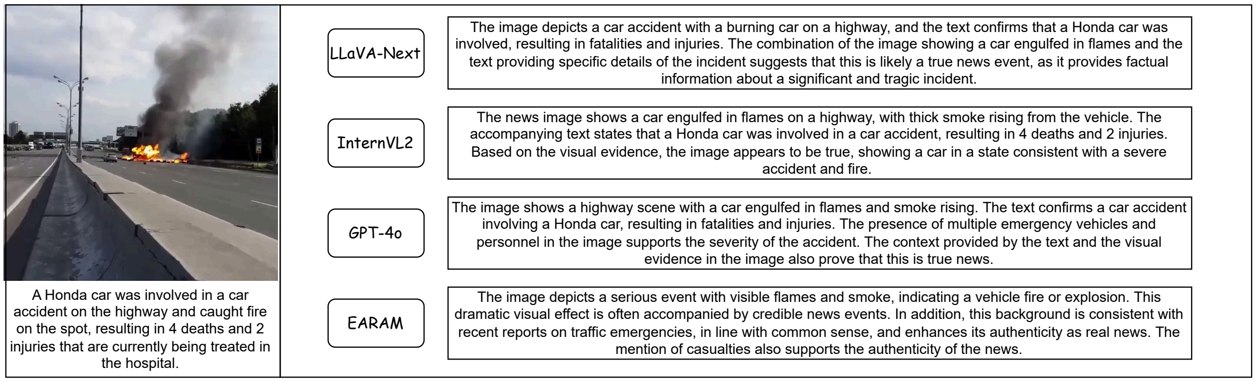
C Examples of Output Explanations for Different Models

Figure 5 illustrates the explanations provided by different models for determining the authenticity of the same news.

EARAM's analysis not only describes the visual elements in the image (such as flames and smoke) but also logically infers the credibility of the news event. It highlights that the event "in line with common sense," pointing out common characteristics of emergency traffic accident reports, which enhances the perceived authenticity. This logical reasoning strengthens the credibility of the text. In contrast, the LLaVA-Next and InternVL2 models mainly focus on the surface-level alignment between the image and text, confirming that both describe a car accident. However, they do not delve into the meaning of the visual elements nor consider broader contextual information. GPT-4o mini performs slightly better than the previous two models, as it mentions the presence of emergency vehicles, but still lacks deep analysis of the visual elements and the use of external knowledge. This indicates that our model benefits from prior common-sense analysis and the complementary logical analysis of image and text. The model synthesizes these analyses from various perspectives, resulting in more detailed and logically coherent explanations.

Additionally, during our experiments, we observed that GPT-4o mini occasionally refuses to answer, stating that it is unable

[3] https://github.com/chroma-core/chroma
[4] https://huggingface.co/BAAI/bge-base-en-v1.5
[5] https://huggingface.co/facebook/metaclip-b16-fullcc2.5b
[6] https://huggingface.co/OFA-Sys/chinese-clip-vit-base-patch16
[7] https://www.deepl.com/en/pro-api



A Honda car was involved in a car accident on the highway and caught fire on the spot, resulting in 4 deaths and 2 injuries that are currently being treated in the hospital.

The image depicts a car accident with a burning car on a highway, and the text confirms that a Honda car was involved, resulting in fatalities and injuries. The combination of the image showing a car engulfed in flames and the text providing specific details of the incident suggests that this is likely a true news event, as it provides factual information about a significant and tragic incident.

The news image shows a car engulfed in flames on a highway, with thick smoke rising from the vehicle. The accompanying text states that a Honda car was involved in a car accident, resulting in 4 deaths and 2 injuries. Based on the visual evidence, the image appears to be true, showing a car in a state consistent with a severe accident and fire.

The image shows a highway scene with a car engulfed in flames and smoke rising. The text confirms a car accident involving a Honda car, resulting in fatalities and injuries. The presence of multiple emergency vehicles and personnel in the image supports the severity of the accident. The context provided by the text and the visual evidence in the image also prove that this is true news.

The image depicts a serious event with visible flames and smoke, indicating a vehicle fire or explosion. This dramatic visual effect is often accompanied by credible news events. In addition, this background is consistent with recent reports on traffic emergencies, in line with common sense, and enhances its authenticity as real news. The mention of casualties also supports the authenticity of the news.

Figure 5: A comparison of the explanation quality generated by different models.

to determine the authenticity of the news based on the available information. Around 30% of its responses were refusals. When such responses occurred, we asked GPT-4o mini to re-answer until a detailed analysis was provided.

D Limitations

We acknowledge that there are certain limitations in our current research. For example, our approach involves multiple calls to LVLMs, which may lead to higher costs compared to traditional methods. However, the cost of using large language models is rapidly decreasing, and we believe this issue will be resolved in the future. Additionally, due to cost and API limitations, as well as the fact that GPT-4o mini frequently refuses to answer questions, stating that it cannot determine the authenticity of the news based on the available knowledge, we opted to use open-source models.

E Discussion and Future Work

One of the main findings in our exploration and experiments is that, in the task of multimodal fake news detection, LVLMs are more suitable for generating multi-perspective analyses to assist in judgment rather than making the final truthfulness determination. This phenomenon is worth further discussion.

It reflects that although LVLMs encompass extensive knowledge and possess a profound understanding of real-world contexts and logical reasoning, they struggle to fully leverage their capabilities for specific downstream tasks. This challenge may not be easily resolved through prompt engineering or fine-tuning.

Thus, combining MSLMs and LVLMs for specific downstream tasks might be a more efficient way to harness LVLMs' internal knowledge. This approach can also be extended to other similar tasks. In the future, we can explore how to integrate MSLM capabilities into LVLMs to fully and efficiently utilize their knowledge across domains.

F Ethical Considerations

The multimodal fake news detection (MFND) system developed in this study aims to curb the spread of misinformation and mitigate its negative impact on society, especially in sensitive areas such as public opinion, social stability, and elections. However, we are also aware that malicious users may attempt to exploit system vulnerabilities and reverse-engineer the MFND to spread false or misleading information. We strongly condemn any such misuse or attempts to circumvent the system and recommend introducing human review mechanisms to enhance the security and accuracy of the system to prevent these risks.

In addition, we fully recognize the potential psychological impact that the task of detecting fake news may have on evaluators. We will inform them in advance about the nature of the work, limit their exposure frequency, encourage timely breaks, and provide access to mental health support resources. The contributions of the evaluators are crucial to this study, and we will provide appropriate compensation to acknowledge their efforts.

Regarding data usage, we are committed to adhering to data privacy and ethical standards. All fake news data have been anonymized and do not contain any personal user information. The collection and use of user data strictly comply with the terms of use and privacy policies of the relevant platforms. All data processing and analysis in this study are conducted in accordance with current laws, regulations, and ethical guidelines, aiming to protect user privacy while enhancing the public's ability to identify misinformation.