LEMMA: Towards LVLM-Enhanced Multimodal Misinformation Detection with External Knowledge Augmentation

Anonymous ACL submission

Abstract

The rise of multimodal misinformation on social platforms poses significant challenges for individuals and societies. Its increased credibility and broader impact make detection more complex, requiring robust reasoning across diverse media types and profound knowledge for accurate verification. The emergence of Large Vision Language Model (LVLM) offers a potential solution to this problem. Leveraging their proficiency in processing visual and textual information, LVLM demonstrates promising capabilities in recognizing complex information and exhibiting strong reasoning skills. We investigate the potential of LVLM on multimodal misinformation detection and find that even though LVLM has a superior performance compared to LLMs, its profound reasoning may present limited power with a lack of evidence. Based on these observations, we propose LEMMA: LVLM-Enhanced Multimodal Misinformation Detection with External Knowledge Augmentation. LEMMA leverages LVLM intuition and reasoning capabilities while augmenting them with external knowledge to enhance the accuracy of misinformation detection. Our external knowledge extraction module adopts multi-query generation and image source tracing to enhance the rigor and comprehensiveness of LVLM's reasoning. We observed that LEMMA improves the accuracy over the top baseline LVLM by **9%** and **13%** on *Twitter* and *Fakeddit* datasets respectively. [1]

1 Introduction

Multimodal misinformation, originating from the integration of multimedia on social platforms, raises significant concerns for individuals and societies. The contents of such misinformation can be readily consumed by the audience, often gaining a higher level of credibility and causing a border impact compared to textual misinformation (Michael Hameleers and Bos, 2020; Zannettou et al., 2018). Unlike unimodal misinformation, detecting multimodal misinformation is more challenging, requiring robust reasoning to decipher cross-modal clues, coupled with the necessity for profound knowledge to verify the factuality of the essential information.

The rise of Large Language Models (LLMs) (Zhao et al., 2023) has significantly reshaped traditional NLP tasks, while recent efforts are leveraging LLMs to combat misinformation (Chen and Shu, 2023; Hu et al., 2023). However, these efforts are limited by LLMs' inability to process non-textual resources. Therefore, the recent emergence of Large Vision Language Models (LVLM) (OpenAI et al., 2023) provides a good opportunity to forward this line of research and here are several intuitions of adopting LVLM into combating multimodal misinformation: Firstly, the pretraining process with large-corpus provides LVLM with a profound understanding of real-world knowledge (Du et al., 2023) so that it has the potential to recognize complex information such as terms or entities appearing in the multimodality. Secondly, LVLM exhibits a strong reasoning capability through showcasing its remarkable performance on various tasks such as arithmetic reasoning (Amini et al., 2019), question answering (Kamalloo et al., 2023), and symbolic reasoning (Wei et al., 2023). Thus, it has the potential to generate strong reasoning from multimodalities even in the zero-shot manner (Kojima et al., 2023). Moreover, LVLM presents a promising capability in incorporating external knowledge by utilizing retrieval-based tools, which has proved to be a beneficial functionality, particularly in tasks that demand fact-checking (Fatahi Bayat et al., 2023).

Considering the aforementioned motivations, our primary objective is to investigate the following research questions: **Can LVLM effectively detect**

[1] The code is available at https://anonymous.4open.science/r/LEMMA

multimodal misinformation given their inherent capabilities? We discover that LVLM can generally demonstrate satisfactory performance with its promising capability to process and reason about complex multimodal content. Despite these advances, current models still struggle when external contextual understanding is necessary for accurate misinformation detection. Traditional approaches to augmenting LLMs with external knowledge and up-to-date information, such as Retrieval-Augmented Generation (RAG) (Lewis et al., 2021), often rely on directly generating queries from factual text. While effective for simple fact-checking, this method falls short in addressing the deceptive nature of multimodal misinformation. (Chen and Shu, 2023). In addition, those methods usually can only capture semantic relevance and are unable to handle logical connections, resulting in information loss.

To bridge this gap, we introduce LEMMA: **L**VLM-**E**nhanced **M**ultimodal **M**isinformation Detection with External Knowledge **A**ugmentation. Unlike conventional methods which usually convert all modalities into textual information for analysis, LEMMA conducts parallel text and image searches to gather comprehensive evidence to enhance the quality of LVLM's reasoning. In addition, our approach utilizes a reasoning-aware multi-query generation that allows the model to evaluate the relevance of details within the broader misinformation context, thereby preventing over-focus on trivial details. What's more, we adopt a coarse to fine-grained distillation module that can effectively improve the quality of retrieval evidence. Our experiments show that LEMMA significantly improves accuracy over the top baseline LVLM by **9%** and **13%** on the *Twitter* and *Fakeddit* datasets. In summary, the major contributions of this paper are as follows:

- We present a comprehensive empirical evaluation of LVLM capabilities on multimodal misinformation detection based on its inherited capability.
- We propose LEMMA, a simple yet effective LVLM-based approach that utilizes the benefits of LVLM intuition and reasoning capability with advanced, reasoning-based query generation and evidence filtering.
- We design an ad-hoc external knowledge extraction module that adopts multi-query generation and image source tracing to enhance the rigor and comprehensiveness of LVLM's reasoning.

2 Related Work

2.1 Multimodal Misinformation Detection

With the proliferation of multimedia resources, multimodal misinformation detection has gained increasing attention in recent years due to its potential threat to the dissemination of genuine information (Alam et al., 2022). To identify multimodal misinformation, a traditional way is to evaluate the consistency between multimodality. To be specific, such evaluation can be realized by approaches such as multimodality feature representation learning (Wang et al., 2018; Shu et al., 2019; Xue et al., 2021), using image captioning model (Zhou et al., 2020) and vision transformer (Ghorbanpour et al., 2021). However, these methods usually rely on a deep learning-based model, which leads to the weakness of interpretability. To address this issue, Liu et al. (2023b) tries to improve interpretability by integrating explainable logic clauses. In addition, Fung et al. (2021) proposes InfoSurgeon which attempts to solve this task by extracting fine-grained information in multimodality. However, this method presents limited precision and recall due to the limitation of automatic IE techniques. Furthermore, these methods suffer from the inherent limitations of the training process, which restrict their generalizability. Therefore, recently researchers have increasingly focused on leveraging LVLMs to tackle multimodal misinformation. After Lyu et al. (2023) illustrates LVLM's effectiveness in the task, key areas of this research extend to developing targeted solutions to combat specific types of multimodal misinformation (Qi et al., 2024), addressing challenges related to domain shift (Liu et al., 2024), and enhancing interpretability (Wang et al., 2024). These studies reflect LVLM as a promising solution to multimodal misinformation detection.

2.2 Retrieval-Augmented Generation (RAG) for LLM/LVLM

RAG is an advanced technique that combines the power of LLM/LVLM with information retrieval techniques. This method was originally designed to address the hallucination issue in text generation by LLMs (Lewis et al., 2020). In addition, RAG approach is frequently applied in tasks requiring factual consistency, such as open-domain question answering (Zhu et al., 2021), fact-checking (Maynez et al., 2020) and code generation (Vaithilingam et al., 2022), which demonstrates its promise. However, traditional RAG suffers from limitations such as static retrieval and lack of efficiency, which prompts researchers to develop more advanced versions to overcome these challenges. For example, Rackauckas (2024) demonstrates combining RAG and reciprocal rank fusion to improve comprehensiveness, Mallen et al. (2023) proposes to evaluate query complexity based on entity frequency and Jeong et al. (2024) incorporates a question complexity classifier to adjust the external knowledge retrieval strategy for question answering. Meanwhile, Merth et al. (2024) introduces superposition prompting to process input documents in parallel and (Jin et al., 2024) improves the RAG efficiency through designing a multilevel dynamic caching system. Despite these advancements, the application of RAG in multimodal misinformation detection poses unique challenges. A critical aspect is the ability to discern and prioritize details that are crucial for identifying rumors while minimizing the retrieval of trivial details. To effectively address this requirement, our method incorporates a reasoning-based query generation approach, which guides the LVLM to focus on analyzing the most pertinent information first, thereby enabling targeted searches for external resources.

3 Preliminary

3.1 Task Definition

In this paper, our objective is to explore an LVLM-based solution for multimodal misinformation detection tasks. Given a post or news report which is formatted as an image-text pair ($\mathcal{I}$, $\mathcal{T}$), we seek to classify it into a candidate label set $\mathcal{Y} = \{\text{NonMisinformation}, \text{Misinformation}\}$ based on two major criteria: **1)** whether there is an information inconsistency between $\mathcal{I}$ and $\mathcal{T}$ and **2)** whether there is a factuality issue in either $\mathcal{I}$ or $\mathcal{T}$.

3.2 Exploration

3.2.1 Evaluation Sets

To assess the performance of LVLM on multimodal misinformation detection based on its inherent capability, we mainly evaluate its performance on two representative datasets in the field and the detailed stats for each dataset are presented in Appendix A.

Twitter (Ma et al., 2017) collects multimedia tweets from Twitter platform. The posts in the dataset contain textual tweets, image/video attachments, and additional social contextual information. For our task, we filtered out only image-text pairs as testing samples.

Fakeddit (Nakamura et al., 2019) is designed for fine-grained fake news detection. The dataset is curated from multiple subreddits of the Reddit platform where each post includes textual sentences, images, and social context information. The 2-way categorization for this dataset establishes whether the news is real or false.

As LVLM doesn't necessitate a training phase, we leverage the testing sets directly from all evaluated datasets. Furthermore, we incorporate preprocessing by filtering out overly short tweets based on text length, as overly short texts are not able to provide sufficient information for inconsistency detection.

3.2.2 Approaches

We mainly exploit two fundamental prompting strategies for testing LVLM inherent capabilities on our task:

Direct: In this method, we operate under the assumption that LVLM functions as an independent misinformation detector. Without applying any preprocessing techniques to image and text resources, we directly prompt LVLM to generate its prediction and then provide reasoning, relying solely on its internal knowledge.

Chain of Thought: The Chain of Thought (CoT) mechanism (Wei et al., 2023) has demonstrated significant enhancement in the ability of LLMs to engage in complex reasoning tasks. Based on the Direct method, we further incorporate the phrase "*Let's think step by step*" after the prompt. And LVLM is asked to first generate its reasoning and finally give out its prediction.

3.2.3 Experiment Settings

We take GPT-4V as a representative model to evaluate LVLM capability on multimodal misinformation detection. In our pursuit to understand the evolution of LVLMs, we also implement the aforementioned prompting approaches with GPT-3.5 and GPT-4. Since these models are not inherently multimodal, we conduct a preprocessing step by converting images into textual summaries to facilitate the input of multimodal content. Additionally, to ensure a more comprehensive evaluation, we in-

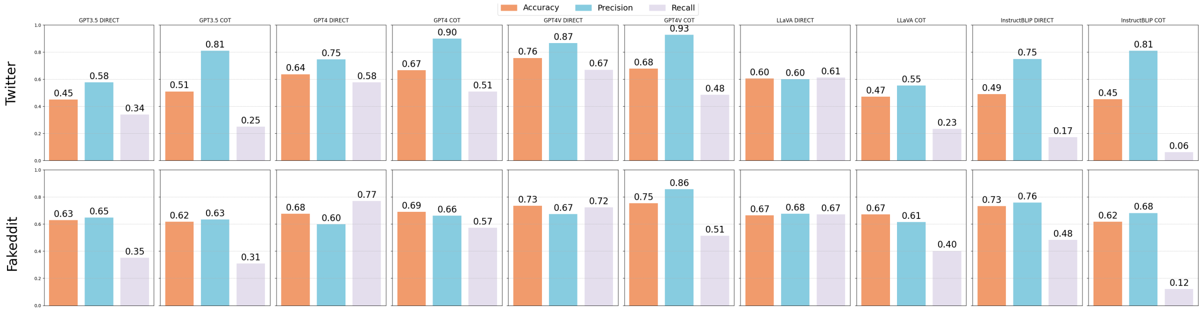
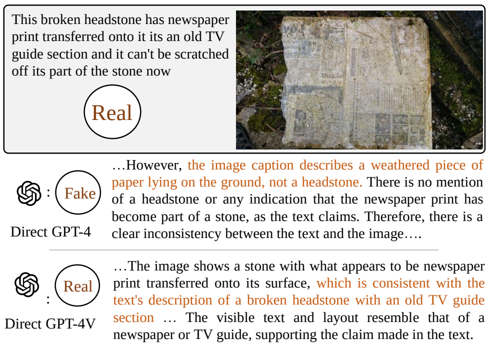


Figure 1: Comparison of performance metrics across various LVLMs/LLMs (GPT-3.5, GPT-4, GPT-4V, LLaVA, and InstructBLIP) and prompting methods (DIRECT and CoT) on two different datasets (*Twitter* and *Fakeddit*).

corporate other two famous LVLMs, LLaVA (Liu et al., 2023a) and InstructBLIP (Dai et al., 2023) into our experiments, which allows us to scrutinize how various LVLMs perform and to identify more general observations.


This broken headstone has newspaper print transferred onto it its an old TV guide section and it can't be scratched off its part of the stone now

…However, the image caption describes a weathered piece of paper lying on the ground, not a headstone. There is no mention of a headstone or any indication that the newspaper print has become part of a stone, as the text claims. Therefore, there is a clear inconsistency between the text and the image….

…The image shows a stone with what appears to be newspaper print transferred onto its surface, which is consistent with the text's description of a broken headstone with an old TV guide section … The visible text and layout resemble that of a newspaper or TV guide, supporting the claim made in the text.

Figure 2: An example of a real *Fakeddit* post where GPT-4V makes a correct prediction based on successfully extracting cross-modal alignment, while GPT-4 fails.

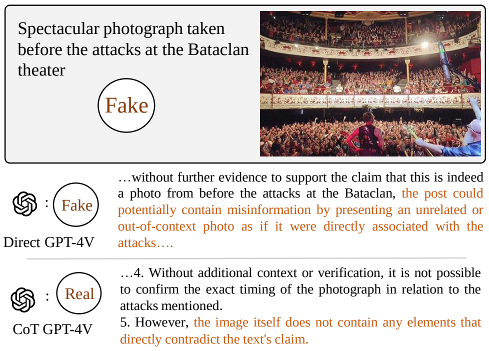
3.2.4 Observation on Preliminary Result

Figure 1 showcases the preliminary result of employing fundamental prompting strategies on two datasets using various LLMs/LVLMs. Upon scrutinizing the predictions and accompanying rationale, we deduce the following insights:

1. **GPT-4V surpasses other LLMs/LVLMs in comprehending cross-modal interaction**: Across both datasets and prompting methods, GPT-4V demonstrates superior performance over other LLMs (like GPT-3.5 and GPT-4) and LVLMs (such as LLaVA and InstructBLIP). This superiority, when compared to LLMs, can be attributed primarily to its enhanced capability for multimodal understanding. For instance, Figure 2 shows a real *Fakeddit* post in which GPT-4V accurately extracts correlations between image and text. However, GPT4 struggles in extracting such correlation which eventually leads to a wrong decision. On the other hand, despite their pretraining for better multimodal capabilities, InstructBLIP and LLaVA tend to underperform due to their failure to follow instructions consistently and the mismatch between training corpus and specific task requirements, which eventually leads to the performance disparity in favor of GPT-4V.

2. **In the absence of external evidence, reasoning-enhanced methods have very limited potential for performance improvement**: While CoT has already demonstrated superior performance in various tasks, its efficacy is limited in multimodal misinformation contexts when used with LVLMs. Specifically, while CoT may increase precision, it consistently yields lower recall compared to the Direct method, which suggests a tendency towards over-conservatism. Considering the importance of real-time information to misinformation detection, such conservative bias likely stems from the inherent limitations in reasoning without adequate supporting evidence, highlighting an essential trade-off between precision and recall in misinformation detection. For instance, Figure2 depicts a fabricated Twitter tweet that requires external evidence for an accurate decision. In such scenarios, CoT tends to guide LVLM towards a conservative stance.

Based on these observations, although LVLM can achieve decent performance based on its inherent capability, it has limited power to make correct judgments when further evidence is necessary for the correct prediction. Therefore, with the insertion of external knowledge, LVLM is expected to achieve better performance.



Spectacular photograph taken before the attacks at the Bataclan theater

…without further evidence to support the claim that this is indeed a photo from before the attacks at the Bataclan, the post could potentially contain misinformation by presenting an unrelated or out-of-context photo as if it were directly associated with the attacks….

…4. Without additional context or verification, it is not possible to confirm the exact timing of the photograph in relation to the attacks mentioned.
5. However, the image itself does not contain any elements that directly contradict the text's claim.

Figure 3: An example of a fabricated *Twitter* tweet that shares subtle discrepancies in two modalities, misleading GPT-4V to answer "presence of misinformation"

4 Methodology

This section introduces the proposed **L**VLM-**E**nhanced **M**ulimodal **M**isinformation Detection with External Knowledge **A**ugmentation (**LEMMA**). The pipeline of LEMMA is illustrated in Figure 4. We first delve into the initial stage inference in Section 4.1. Subsequently, we elucidate how we generate reasoning-aware queries to retrieve relevant multimodal evidence from the Internet in Section 4.2. Additionally, we present the methodology for filtering qualified evidence from search results in Section 4.3. Finally, we demonstrate how LEMMA utilizes additional references to refine its final prediction in Section 4.4. The detailed prompt design for each module is shown in Appendix B.3.

4.1 Initial Stage Inference

In the initial phase, LVLM assesses whether posts inherently contain misinformation based on observed cross-modal inconsistencies, and determines whether external information is necessary to make a final judgment. Upon receiving an image-text pair ($\mathcal{I}$, $\mathcal{T}$), LVLM generates an initial prediction $\mathcal{Y}_\mathrm{D}$ and accompanying rationale $\mathcal{R}_\mathrm{D}$ which includes the assessment of consistency level between $\mathcal{I}$ and $\mathcal{T}$. Subsequently, leveraging reasoning $\mathcal{R}_\mathrm{D}$, LVLM is able to autonomously evaluate the necessity for external knowledge based on whether the within-context information is sufficient to conclude the judgment and whether any contents need to be verified. Following this evaluation, LVLM will finalize its decision as the direct prediction if the current information is deemed sufficiently comprehensive. Otherwise, LVLM proceeds to extract external evidence for further analysis to avoid an overly conservative bias. Furthermore, if LVLM thinks the external knowledge is still insufficient for judgment, it will classify this post as "Unverified" in the refined prediction phrase and choose direct prediction instead as the final output. More details in Section 4.4.

4.2 Multimodal Retrieval

In addressing the challenge of potentially conservative bias due to insufficient evidence, we proposed a multimodal retrieval framework that combines reasoning-aware multi-query-based text retrieval and image context retrieval.

4.2.1 Reasoning Aware Multi-Query Retrieval

Traditional retrieval methods often directly use original posts for query construction, leading to potential losses in semantic integrity and difficulties in matching dispersed information (Mallen et al., 2023; Shi et al., 2023). To address these, we employ LVLM to generate multi-faceted queries based on the direct reasoning $\mathcal{R}_\mathrm{D}$. Specifically, LVLM receives the image-text pair ($\mathcal{I}$, $\mathcal{T}$), along with the initial prediction $\mathcal{Y}_\mathrm{D}$ and the reasoning $\mathcal{R}_\mathrm{D}$ generated during the initial stage inference. LVLM first synthesizes a concise title $\mathcal{Q}_\mathrm{t}$ for the post, where a "fake news" prefix is added to increase the likelihood of retrieving content that directly refutes the claims made in $\mathcal{T}$. Then, it reviews direct reasoning $\mathcal{R}_\mathrm{D}$ that identifies the key discrepancies and statements that would suggest potential misinformation and raises several questions $\mathcal{Q}_\mathrm{q}$ to verify them. This ensures that the system prioritizes areas most susceptible to misinformation.

The combined query set ($\mathcal{Q}_\mathrm{t}$, $\mathcal{Q}_\mathrm{q}$), is used to search via the DuckDuckGo Search API (DuckDuckGo, 2023), aiming to retrieve highly relevant documents set $\mathcal{D}$, each annotated with a web title and brief description.

4.2.2 Image Context Retrieval

Traditional retrieval methods often transform multimodal content into textual representations to facilitate analysis, focusing mainly on contextual comprehension. However, as depicted in Figure 5, this approach may overlook essential aspects of misinformation. For example, an image purported to show recent environmental benefits from new solar panels in Germany is actually an old promotional image. To address such discrepancies, image

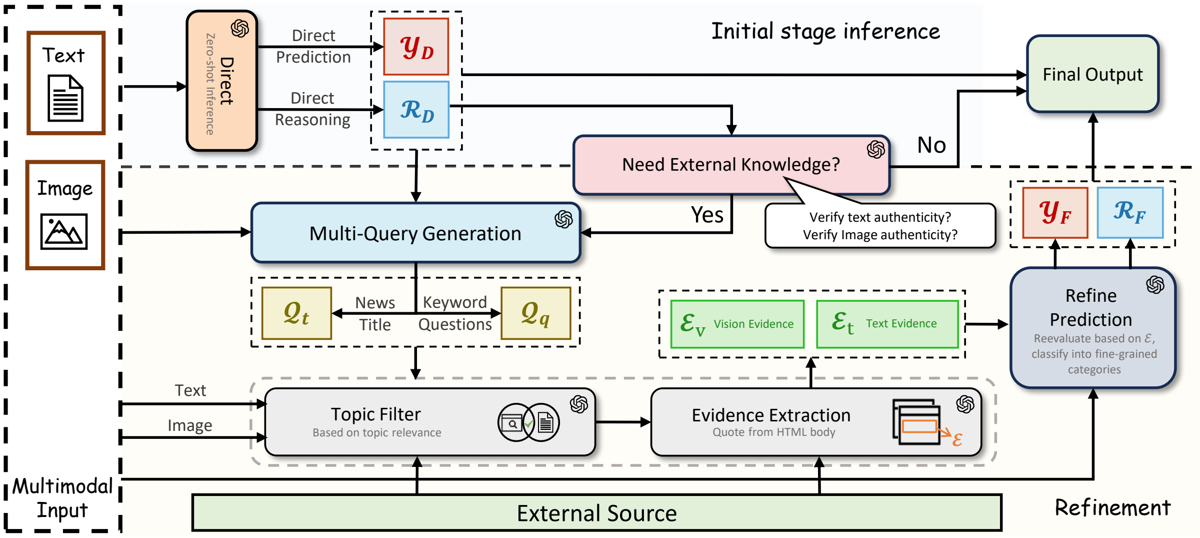


Figure 4: The pipeline of the proposed method (**LEMMA**). The process hinges on two key inputs: multimodal data and selectively filtered evidence gathered from external sources. Components marked with the OpenAI LOGO are developed using the LVLM (GPT-4V).

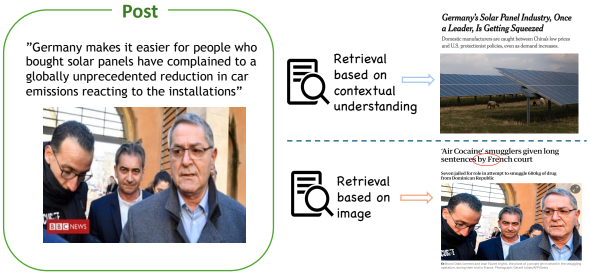

Post
"Germany makes it easier for people who bought solar panels have complained to a globally unprecedented reduction in car emissions reacting to the installations"
Retrieval based on contextual understanding
Germany's Solar Panel Industry, Once a Leader, Is Getting Squeezed
Domestic manufacturers are caught between China's low prices and U.S. protectionist policies, even as demand increases.
Retrieval based on image
'Air Cocaine' smugglers given long sentences by French court
Seven jailed for role in attempt to smuggle 680kg of drug from Dominican Republic
Bruno Odos (centre) and Jean Fauret (right), the pilots of a private jet involved in the smuggling operation, during their trial in France. Photograph: Gérard Julien/AFP/Getty

Figure 5: A fake news example that image retrieval exposes as a reused promotional image

context retrieval technique provides a substantial improvement. By tracing the origin of an image, visual search adds a layer of context that significantly enhances the accuracy of misinformation detection.

To implement the image context retrieval, we utilize the Google search engine to trace the sources of image $\mathcal{I}$ and the exact match technique to pinpoint the sources which contain the pictures that are identical or highly similar to $\mathcal{I}$. Eventually, a list of web page's title is returned as evidence $\mathcal{E}_\text{v}$, which can offer a more accurate estimation of the image's context for later evaluation.

4.3 Resource Distillation

To address the challenge of off-topic or irrelevant information retrieved by search engines, we employ a resource distillation process, refining the traditional chunking technique based on the vector space model which lacks awareness of logical text connections (Lewis et al., 2020). We adopt a coarse to fine-grained distillation approach, similar to LongLLMLingua (Jiang et al., 2023)

4.3.1 Topic Filtering

Initially, the top k relevant resources form a root document set $\mathcal{D}$. The LVLM then evaluates the topic relevance level of each document in $\mathcal{D}$ based on query ($\mathcal{Q}_\text{t}$, $\mathcal{Q}_\text{q}$) and original context $\mathcal{I}$. Eventually, a further refined set $\mathcal{D}'$ is returned, containing only documents that are highly relevant to the post's content. To ensure efficiency, we ask LLM to process a batch of resources in one request.

4.3.2 Evidence Extraction

For each document in $\mathcal{D}'$, we extract the main content along with the publication date. Subsequently, the LLM identifies key segments $\mathcal{S}_\text{i}$ that either support or refute the original post $\mathcal{T}$. The LLM is instructed to extract these segments directly from the HTML body of the document, ensuring they are succinct yet comprehensive, capturing all relevant information. These segments, along with the document's web title and publication date, are then compiled into an evidence entry, formatted as a triplet. The aggregated evidence, $\mathcal{E}_\text{t}$, is a collection of these triplets, forming a structured dataset ready for analysis.

4.4 Refined Prediction

With the set of extracted evidence ($\mathcal{E}_\text{t}$, $\mathcal{E}_\text{v}$) collected from external sources, the model gains a more com-

Dataset	Method	Accuracy	Rumor			Non-Rumor		
			Precision	Recall	F1	Precision	Recall	F1
Twitter	Direct (LLaVA)	0.605	0.688	0.590	0.635	0.522	0.626	0.569
	CoT (LLaVA)	0.468	0.563	0.231	0.635	0.441	0.765	0.560
	Direct (InstructBLIP)	0.494	0.751	0.171	0.277	0.443	0.902	0.599
	CoT (InstructBLIP)	0.455	0.813	0.067	0.112	0.428	0.921	0.596
	Direct (GPT-4)	0.637	0.747	0.578	0.651	0.529	0.421	0.469
	CoT (GPT-4)	0.667	0.899	0.508	0.649	0.545	0.911	0.682
	FacTool (GPT-4)	0.548	0.585	**0.857**	0.696	0.273	0.082	0.125
	Direct (GPT-4V)	0.757	0.866	0.670	0.756	0.673	0.867	0.758
	CoT (GPT-4V)	0.678	0.927	0.485	0.637	0.567	0.946	0.709
	LEMMA	**0.824**	0.943	0.741	**0.830**	**0.721**	0.937	**0.816**
	w/o *initial-stage infer*	0.809	0.932	0.736	0.823	0.699	0.919	0.794
	w/o *visual retrieval*	0.781	**0.953**	0.672	0.788	0.652	**0.949**	0.773
Fakeddit	Direct (LLaVA)	0.663	0.588	0.797	0.677	0.777	0.558	0.649
	CoT (LLaVA)	0.673	0.612	0.400	0.484	0.694	0.843	0.761
	Direct (InstructBLIP)	0.726	0.760	0.489	0.595	0.715	0.892	0.793
	CoT (InstructBLIP)	0.610	0.685	0.190	0.202	0.604	0.901	0.742
	Direct (GPT-4)	0.677	0.598	0.771	0.674	0.776	0.606	0.680
	CoT (GPT-4)	0.691	0.662	0.573	0.614	0.708	0.779	0.742
	FacTool (GPT-4)	0.506	0.476	**0.834**	0.606	0.624	0.232	0.339
	Direct (GPT-4V)	0.734	0.673	0.723	0.697	0.771	0.742	0.764
	CoT (GPT-4V)	0.754	0.858	0.513	0.642	0.720	**0.937**	0.814
	LEMMA	**0.828**	**0.881**	0.706	**0.784**	**0.800**	0.925	**0.857**
	w/o *initial-stage infer*	0.803	0.857	0.692	0.766	0.786	0.891	0.830
	w/o *visual retrieval*	0.792	0.818	0.675	0.740	0.778	0.883	0.854

Table 1: Performance comparison of baseline methods and LEMMA on *Twitter* and *Fakeddit* dataset. We show the result of eight different baseline methods. Additionally, we present the results of two ablation studies: one without initial-stage inference, and the other without resource distillation and evidence extraction. The best two results are **bolded** and underlined.

prehensive understanding of the multimodal content, enabling it to make a more accurate prediction. In detail, the image-text pair $(\mathcal{I}, \mathcal{T})$ is re-introduced to the LVLM, accompanied with the evidence set $(\mathcal{E}_\mathrm{t}, \mathcal{E}_\mathrm{v})$. LVLM is tasked with reevaluating its decision in light of the extracted evidence. Inspired by the fine-grained definition of multimodal misinformation (Nakamura et al., 2019), LVLM is asked to categorize the post into one of six categories: 1) True, 2) Satire, 3) Misleading Content, 4) False Connection, 5) Manipulated Content, or 6) Unverified Content. Categories 2 through 6 correspond to different types of misinformation, while Category 1 indicates real news. LVLM retains its inference from the initial stage if it classifies the post as Category 6, prioritizing conservatism over a potentially risky choice.

5 Experiments

5.1 Experiment Settings

We evaluate LEMMA by comparing it with the following baseline models and methods: **1) LLaVA:** We evaluate LLaVA-1.5-13B (Liu et al., 2023a), which is a state-of-the-art LVLM based on vision instruction tuning, by employing the Direct approach. **2) InstructBLIP:** We evaluate the InstructBLIP (Dai et al., 2023), which is a multimodal transformer designed to perform image-text tasks by leveraging instruction-based finetuning. **3) GPT-4 with Image Summarization:** We evaluate the effectiveness of the fundamental GPT-4 model (without visual understanding). To provide visual context, we construct a GPT4-V-based Image Summarization module, which generates comprehensive textual descriptions corresponding to images. As elaborated in Section 3.2, we employ both the Direct and CoT approaches within this experimental framework. **4) GPT-4 with Factool:**We evaluated FacTool (Chern et al., 2023) with GPT-4 and image summarization as its foundation. Factool is an LLM-based framework that can detect factual inaccuracies within texts. Similar to LEMMA, FacTool incorporates query generation and evidence retrieval to verify claims. However, its methodology is specifically tailored for AI-generated text and does not take LVLM as a backbone. **5) GPT-4V:** We evaluate GPT-4V, also employing the Direct and CoT approaches.

Datasets: We evaluate LEMMA and all the baselines on the *Twitter* and the *Fakeddit* datasets, as introduced in 3.2.

5.2 Performance Comparison

The results presented in Table 1 demonstrate that our proposed LEMMA framework consistently surpasses baseline models on the *Twitter* and *Fakeddit* datasets in terms of both Accuracy and F1 Score. Specifically, LEMMA shows an improvement of approximately 6.7% in accuracy on *Twitter* and a notable 7.4% increase on *Fakeddit* when compared to the best-performing baseline. Compared to FacTool which suffers from an overemphasis on trivial details that makes its predictions overly sensitive, our approach excels in balancing precision and recall, achieving high scores in both metrics. This suggests that LEMMA is effective in minimizing both false positives and false negatives, enhancing the overall quality of its predictions. Additionally, LEMMA demonstrates robust performance across different datasets, confirming its reliability and effectiveness in diverse contexts, essential for practical applications.

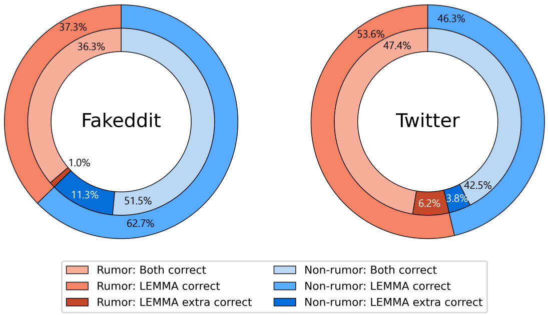


Figure 6: Comparison of the distribution of correct predictions between LEMMA and baseline (GPT-4V).

5.3 Ablation Study

We conduct an ablation study on two modules in LEMMA, with the results shown in Table 1. (*i*) *Initial-stage inference.* We test bypassing LVLM's self-evaluation of external evidence necessity, forcing it to search for external evidence for all posts. This led to a 1.5% lower accuracy on Twitter and a 2.5% decrease on Fakeddit compared to the original version. We hypothesize that this is because LEMMA may be overly sensitive to the subtle differences between the external evidence and the original post. (*ii*) *Visual Retrieval.* We also implement a version without visual context retrieval, resulting in a 0.8% drop in accuracy on Twitter and a 0.6% drop on Fakeddit, suggesting that the image sources provide valuable context, informing LVLM of the true significance behind the visual input, thereby enhancing the overall reasoning quality.

5.4 Result Analysis

We conduct a statistical analysis to compare the accuracy distribution between LEMMA and Direct (GPT-4V). From Figure 6, we have the following observations: First, we observe that LEMMA accurately replicates over 98% of Direct (GPT-4V) correct predictions in *Fakeddit*, while in *Twitter*, this figure stands at over 96%. This suggests that LEMMA maintains an advantage in retaining the inherent capabilities of GPT-4V. Furthermore, in *Fakeddit* and *Twitter*, LEMMA exhibits approximately 13% and 9% additional gains relative to Direct (GPT-4V). Such performance advantages can be attributed to external knowledge providing LEMMA with more evidence favorable for inference, thereby making its reasoning performance more robust.

6 Conclusion

In this study, we investigated the capability of LVLMs in multimodal misinformation detection and discovered the significant importance of providing external information to enhance LVLM performance. Then we proposed LEMMA, a framework designed to enhance LVLMs by utilizing a reasoning-aware query set for effective multimodal retrieval and by integrating external knowledge sources. Our experiments on the Twitter and Fakeddit datasets demonstrated that LEMMA significantly outperforms the top baseline LVLM, achieving accuracy improvements of **9%** and **13%**, respectively. While there is room for further refinement of knowledge source interfaces and filtering, we believe LEMMA is an extensible approach applicable to interpretability-critical reasoning tasks at the intersection of vision, language, and verification.

7 Limitations

We recognize several limitations. 1) Due to the integration of external knowledge sources and multiple

LVLM-based modules, the LEMMA framework may suffer from increased computational complexity and latency. This setup can hinder its scalability and efficiency, particularly in real-time environments where rapid processing is crucial. 2) Our study did not thoroughly examine LEMMA's sensitivity to different prompts. Given the constraints of our study, we defer the exploration of prompt sensitivity to future experiments. 3) The Evaluation datasets are limited to short social media posts due to dataset availability constraints, leaving LEMMA's performance on longer texts untested.

8 Ethics Statement

We acknowledge that our work is aligned with the ACL Code of the Ethics [2] and will not raise ethical concerns. We do not use sensitive datasets/models that may cause any potential issues/risks.

References

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. 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)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of llms: Opportunities and challenges.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative ai – a tool augmented framework for multi-task and multi-domain scenarios.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. 2023. Guiding pretraining in reinforcement learning with large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8657–8677. PMLR.

DuckDuckGo. 2023. Duckduckgo.

Farima Fatahi Bayat, Kun Qian, Benjamin Han, Yisi Sang, Anton Belyy, Samira Khorshidi, Fei Wu, Ihab Ilyas, and Yunyao Li. 2023. FLEEK: Factual error detection and correction with evidence retrieved from external knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 124–130, Singapore. Association for Computational Linguistics.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil. 2021. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698, Online. Association for Computational Linguistics.

Faeze Ghorbanpour, Maryam Ramezani, Mohammad A. Fazli, and Hamid R. Rabiee. 2021. FNR: A similarity and transformer-based approach to detect multi-modal fake news in social media. *CoRR*, abs/2112.01131.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2023. Bad actor, good advisor: Exploring the role of large language models in fake news detection.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression.

Chao Jin, Zili Zhang, Xuanlin Jiang, Fangyue Liu, Xin Liu, Xuanzhe Liu, and Xin Jin. 2024. Ragcache: Efficient knowledge caching for retrieval-augmented generation.

Ehsan Kamalloo, Nouha Dziri, Charles L. A. Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

[2] https://www.aclweb.org/portal/content/ acl-code-ethics

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning.

Hui Liu, Wenya Wang, and Haoliang Li. 2023b. Interpretable multimodal misinformation detection with logic reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9781–9796, Toronto, Canada. Association for Computational Linguistics.

Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng He. 2024. Fakenewsgpt4: Advancing multimodal fake news detection through knowledge-augmented lvlms.

Hanjia Lyu, Jinfa Huang, Daoan Zhang, Yongsheng Yu, Xinyi Mou, Jinsheng Pan, Zhengyuan Yang, Zhongyu Wei, and Jiebo Luo. 2023. Gpt-4v (ision) as a social media analysis engine. *arXiv preprint arXiv:2311.07547*.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717, Vancouver, Canada. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization.

Thomas Merth, Qichen Fu, Mohammad Rastegari, and Mahyar Najibi. 2024. Superposition prompting: Improving and accelerating retrieval-augmented generation.

Toni G.L.A. Van Der Meer Michael Hameleers, Thomas E. Powell and Lieke Bos. 2020. A picture paints a thousand lies? the effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication*, 37(2):281–301.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2019. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection.

Zackary Rackauckas. 2024. Rag-fusion: A new take on retrieval augmented generation. *International Journal on Natural Language Computing*, 13(1):37–47.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 395–405, New York, NY, USA. Association for Computing Machinery.

Priyan Vaithilingam, Tianyi Zhang, and Elena L. Glassman. 2022. Expectation vs.nbsp;experience: Evaluating the usability of code generation tools powered by large language models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA. Association for Computing Machinery.

Longzheng Wang, Xiaohan Xu, Lei Zhang, Jiarui Lu, Yongxiu Xu, Hongbo Xu, Minghao Tang, and Chuang Zhang. 2024. Mmidr: Teaching large language model to interpret multimodal misinformation via knowledge distillation.

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*, KDD '18, page 849–857, New York, NY, USA. Association for Computing Machinery.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 58(5):102610.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities.

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.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Safe: Similarity-aware multi-modal fake news detection.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization.

A Dataset Statistics

Table 2 shows the detailed statistics of two datasets for the testing. We filter the original test sets of the two datasets to exclude overly short texts because they often lack sufficient contextual details. This means overly short texts are not good test cases to determine LVLM's capability for multimodal misinformation detection.

Dataset	Num Rumor	Num Non-rumor	Language Distribution
Twitter	448	321	English 78%, French 9%, Spanish 4%, Other 9%
Fakeddit	342	464	English 99%, Other 1%

Table 2: Dataset Statistics

B Experiment Prompt Template

This section shows the templates for each prompting method that we have examined in section 5.2:

B.1 Direct Prompting

As shown in Figure 7, this method involves directly prompting the model with a text-image pair to determine the presence of misinformation. It must be noted that the rules we provided are designed to streamline the assessment process, ensuring that the model controls the output format while focusing on key indicators of misinformation.

```
You are given a piece of **Input Text** and an image. Your task is to predict whether misinformation is present. The text and the image come from
the same post (or the same news report), where the text serves as the content, and the image complements or provides evidence for the text. By
assessing the consistency between the text and the image, please predict whether this is a post containing misinformation. Please follow the Rules
below:

Rules:
Generate a JSON object with three properties: 'label', 'explanation' and 'external knowledge'.
The return value of 'label' property should be selected from ["real", "fake"].
real indicates that no misinformation is detected.
fake indicates that misinformation is detected.
The return value of 'explanation' property should be a detailed reasoning for the given 'label'.

Note that your response will be passed to the python interpreter, SO NO OTHER WORDS! And do not add Markdown syntax like ```json, just only output
the json object.

Input Text:

{TEXT}

Your Response:
```

Figure 7: Prompt Template for Direct Approach

B.2 CoT Prompting

As illustrated in Figure 8, CoT extends the assessment process by incorporating a more explicit and detailed reasoning pathway. Similar to the B.1, the output format remains controlled.

```
You are given a piece of **Input Text** and an image. Your task is to predict whether misinformation is present. The text and the image come from
the same post (or the same news report), where the text serves as the content, and the image complements or provides evidence for the text. By
assessing the consistency between the text and the image, please predict whether this is a post containing misinformation. Please follow the Rules
below:

# Rules:
1. Start your reasoning with an evaluation based on the sentence 'Let's think step by step'.
2. Output your complete reasoning in the subsequent lines.
3. In the final line, use a single word to indicate whether misinformation exists, which should be selected from ["Real", "Fake"].
"Real" indicates that no misinformation is detected.
"Fake" indicates that misinformation is detected.

# Input Text:

{TEXT}

# Your Response:
Let's think step by step,
```

Figure 8: Prompt Template for CoT Approach

B.3 LEMMA Prompting

B.3.1 Initial Stage Inference

Shown in Figure 9, The prompt for this stage resembles the one from B.1, with the addition of an example to preserve the output format. This is crucial for deriving the reasoning needed for subsequent steps.

```
You are given a piece of **Input Text** and an image. Your task is to predict whether misinformation is present. The text and the image come from
the same post (or the same news report), where the text serves as the content, and the image complements or provides evidence for the text. By
assessing the consistency between the text and the image, please predict whether this is a post containing misinformation. Please follow the Rules
below:

Rules:
Generate a JSON object with three properties: 'label', 'explanation' and 'external knowledge'.
The return value of 'label' property should be selected from ["real", "fake"].
real indicates that no misinformation is detected.
fake indicates that misinformation is detected.
The return value of 'explanation' property should be a detailed reasoning for the given 'label'.

Note that your response will be passed to the python interpreter, SO NO OTHER WORDS! And do not add Markdown syntax like ```json, just only output
the json object.

Example output (JSON):
{{
   "label": "real",
   "explanation": "The image shows a concert venue filled with people who appear to be enjoying a performance, which is consistent with the text's
description of a photo taken at the start of a concert in Paris. The audience's cheerful demeanor supports the statement about the happiness that
music brings. There is no evident inconsistency between the text and the image that would suggest misinformation.",
}}

Input Text:

{TEXT}

Your Response:
```

Figure 9: Instruction for initial stage inference

```
You are given a piece of **Reasoning** about the first-stage decision of the authenticity of multimedia news post, the **Text** is the text part
of the post and you have already got the image part of the post. Your task is to decide whether additional evidence is needed for predicting
whether misinformation is present.

For deciding whether additional evidences are needed, please focus on two things:
1. Whether the authenticity of events is suspicious.
2. Whether the authenticity of the image is suspicious.

Note that you should not easily judge that one post is "true", normally you need more external resources.

You should only respond in format as described below. DO NOT RETURN ANYTHING ELSE. START YOUR RESPONSE WITH '{{'.
[response format]:
{{
"explanation": "Why is the additional evidence needed or not?"
"external knowledge": "Yes" if you think additional evidence is needed, "No" otherwise.
}}


Input Reasoning:

{REASONING}

Input Text:

{TEXT}

Your Response:
```

Figure 10: Instruction for External Knowledge

B.3.2 Necessity of External Knowledge

Based on the initial stage inference in Section B.3.1, LEMMA evaluates the necessity of incorporating external knowledge. This assessment is guided by rules specified in the prompt, which scrutinize the reasoning derived during the first stage. The decision to proceed to subsequent stages is contingent on whether the direct reasoning suggests the need for external verification to support or refute the findings. The detailed procedure of this evaluation is depicted in Figure 10.

B.3.3 Reasoning Aware Multi-query Generation

At this stage, we input both the original image-text pair and the reasoning derived from B.3.1 to generate the following queries: **1)** a title for the post. **2)** two questions related to contents that need to be verified. The design of the prompt is shown in Figure 11

B.3.4 Topic Relevance Filter

When we obtained the resources from the search engine (We use title and queries derived from B.3.2 to search resources), we use the following prompt to check whether each search result is related to the

````
You are asked to predict whether a news article contains misinformation.

The text of this news is:

{TEXT}

Your original prediction is {PREDICTION}. (0 for no misinformation, 1 for presence of misinformation, 'None' for no original prediction)

Your original reasoning based on that prediction is:

{REASONING}

External sources can better help you make the judgment. Please come up with a title for this news first, then list two questions/phrases/sentences
that you would like to search on a public search engine, such as Google. Carefully design your question so that it can return the most helpful
results for making your final prediction and reasoning. Please use English to generate your title and questions.

Text Input example 1:

'Is baltimore's prosecutor wrong about freddie grays legal knife the weapon police described is definitely illegal so why did marilyn mosby say it
wasn't the answer hinges on a single spring'

Output example (JSON) 1:
{{
   "title":"Freddie Gray's Knife: Legal or Not?",
   "questions":[
       "Was Freddie Gray's knife legal?",
       "Marilyn Mosby's comments on Freddie Gray's Legal Knife"
   ]
}}

Text Input example 2:

'RT @danrem: Konon, inside The Bataclan concert before the attack. How life can change in a second. #Pray4Paris'

Output example (JSON) 2:
{{
   "title":"Inside The Bataclan Concert Moments Before the Attack",
   "questions":[
       "the full story of what happened in the Bataclan | Paris attacks",
       "Authenticity of images from Bataclan before the attack"
   ]
}}

Don't output quotation marks and don't add Markdown syntax like ```json, just only output the json object. Your response:
````

Figure 11: Instruction for Query Generation

context based on the queries and title we derived in B.3.2. Each resource will be labeled as True if found relevant, otherwise False. To ensure the efficiency, LVLM is asked to process a batch of resources in one request, using JSON to manage the output format. The design of the prompt is shown in Figure 12

````
Your task is to filter the off-topic search result. You will be provided a piece of text. You have to determine the topic of the text. Then, you
will be provided the search result in JSON format. For each entry, there is a unique integer key serving as the id of each entry. The value of
each entry consists of three attributes: title, body, url. And  You have to filter the off-topic search result according to the content of the
title and body. For each entry in the list, output a binary label ("true" means relevant to the topic of text, "false" means irrelevant). Put all
the labels in a JSON dict

Example output format:

{{"0":true, "1":false, "2":false, "3":true, "4":false, "5":true, "6":false, "7":true}}

Text input that you are going to determine the topic:

{TEXT}

Search result in JSON format:

{SEARCH_RESULT}

Your answer (don't include the Markdown syntax like ```json. just directly output the JSON list object. Don't output anything else):
````

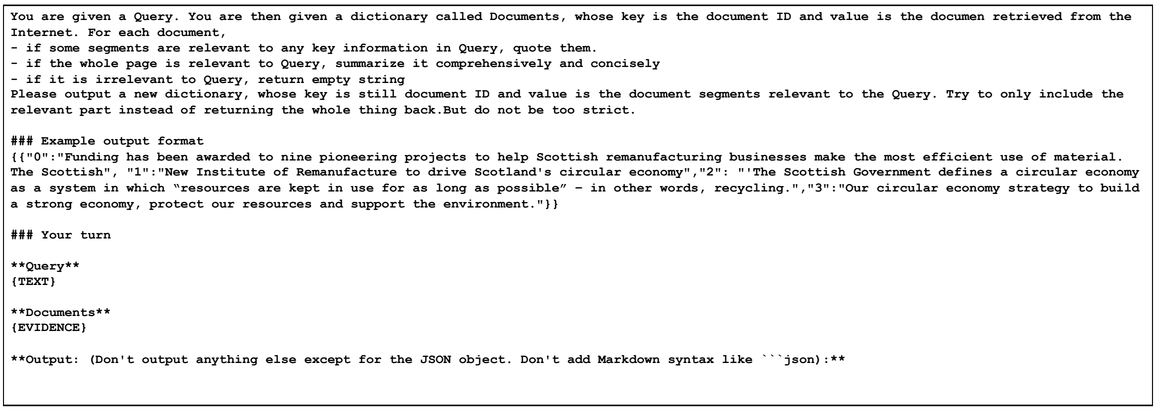
Figure 12: Instruction for Topic Relevance Filter

B.3.5 Evidence Extraction

This stage conducts evidence extraction by quoting or summarizing (if most of the post is relevant) the contents from remaining resources in the last stage. The model is asked to keep the extracted evidence concise, while avoiding excessive strictness. The design of the prompt is shown in Figure 13

B.3.6 Refined Prediction

Upon completing evidence extraction, the model reevaluates the image-text pair post, incorporating evidences retrieved from both text search and image search. Additionally, a fine-grained definition of misinformation (The detail is presented as Appendix C). is utilized for this reassessment. The design of the prompt is shown in Figure 14



```
You are given a Query. You are then given a dictionary called Documents, whose key is the document ID and value is the documen retrieved from the
Internet. For each document,
- if some segments are relevant to any key information in Query, quote them.
- if the whole page is relevant to Query, summarize it comprehensively and concisely
- if it is irrelevant to Query, return empty string
Please output a new dictionary, whose key is still document ID and value is the document segments relevant to the Query. Try to only include the
relevant part instead of returning the whole thing back.But do not be too strict.

### Example output format
{{"0":"Funding has been awarded to nine pioneering projects to help Scottish remanufacturing businesses make the most efficient use of material.
The Scottish", "1":"New Institute of Remanufacture to drive Scotland's circular economy","2": "'The Scottish Government defines a circular economy
as a system in which “resources are kept in use for as long as possible” – in other words, recycling.","3":"Our circular economy strategy to build
a strong economy, protect our resources and support the environment."}}

### Your turn

**Query**
{TEXT}

**Documents**
{EVIDENCE}

**Output: (Don't output anything else except for the JSON object. Don't add Markdown syntax like ```json):**
```

Figure 13: Instruction for Evidence Extraction

C Fine-grained definition of misinformation

This section illustrates the fine-grained definition of misinformation which is used in Refined Prediction stage:

1. **True**: True content is accurate in accordance with fact. Eight of the subreddits fall into this category, such as usnews and mildly interesting. The former consists of posts from various news sites. The latter encompasses real photos with accurate captions.

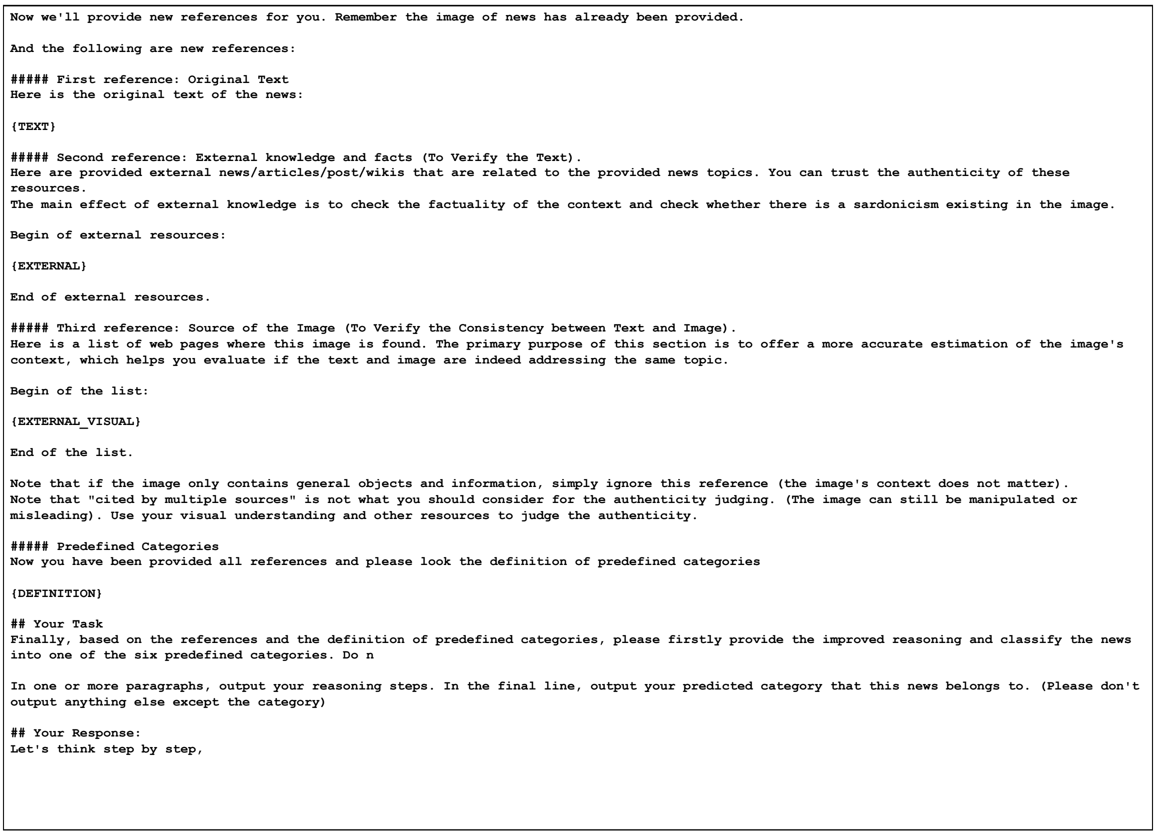
2. **Satire/Parody**: This category includes content that presents true contemporary information in a satirical or humorous manner, often leading to its misinterpretation as false. Examples can be found on platforms like Reddit's T̈he Onion,f̈eaturing headlines such as M̈an Lowers Carbon Footprint By Bringing Reusable Bags Every Time He Buys Gas.,̈ Satire that clearly identifies itself as such and is intended purely for entertainment or social commentary purposes should not be considered misinformation.

3. **Misleading Content**: This category comprises information deliberately manipulated to deceive the audience.

4. **False Connection**: This category encompasses instances where there is a disconnect between the information conveyed by an image and the essential details provided in the accompanying text. It may involve situations where the event depicted in the image does not align with or contradicts the narrative described in the text, leading to potential misinterpretation or misunderstanding.

5. **Manipulated Content**: This category consists of content that has been intentionally altered through manual photo editing or other forms of manipulation. For instance, comments on platforms like the p̈hotoshopbattless̈ubreddit often contain doctored versions of images submitted to the platform.

6. **Unverified**: This category includes news or content for which the presence of misinformation cannot be definitively determined based solely on the available evidence. Additional evidence or verification may be required to confirm or refute the accuracy of the information. This category encompasses situations where there is insufficient information or conflicting sources to make a conclusive determination regarding the accuracy of the content.



```
Now we'll provide new references for you. Remember the image of news has already been provided.

And the following are new references:

##### First reference: Original Text
Here is the original text of the news:

{TEXT}

##### Second reference: External knowledge and facts (To Verify the Text).
Here are provided external news/articles/post/wikis that are related to the provided news topics. You can trust the authenticity of these
resources.
The main effect of external knowledge is to check the factuality of the context and check whether there is a sardonicism existing in the image.

Begin of external resources:

{EXTERNAL}

End of external resources.

##### Third reference: Source of the Image (To Verify the Consistency between Text and Image).
Here is a list of web pages where this image is found. The primary purpose of this section is to offer a more accurate estimation of the image's
context, which helps you evaluate if the text and image are indeed addressing the same topic.

Begin of the list:

{EXTERNAL_VISUAL}

End of the list.

Note that if the image only contains general objects and information, simply ignore this reference (the image's context does not matter).
Note that "cited by multiple sources" is not what you should consider for the authenticity judging. (The image can still be manipulated or
misleading). Use your visual understanding and other resources to judge the authenticity.

##### Predefined Categories
Now you have been provided all references and please look the definition of predefined categories

{DEFINITION}

## Your Task
Finally, based on the references and the definition of predefined categories, please firstly provide the improved reasoning and classify the news
into one of the six predefined categories. Do n

In one or more paragraphs, output your reasoning steps. In the final line, output your predicted category that this news belongs to. (Please don't
output anything else except the category)

## Your Response:
Let's think step by step,
```

Figure 14: Instruction for Refined Prediction