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 compared to textual misinformation make detection 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. In this paper, we first investigate the potential of LVLM on multimodal misinformation detection. We 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 method improves the accuracy over the top baseline LVLM by **7%** and **13%** on *Twitter* and *Fakeddit* datasets respectively.

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). In contrast to the misinformation within unimodal contexts, detecting multimodal misinformation presents a more challenging task, which is attributed to the inherent need for robust reasoning capabilities to analyze cross-modal clues, coupled with the necessity for profound knowledge to verify the factuality of the 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, their attempts in such a direction have been hindered by the limitation that LLMs could only process 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 is 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 multimodal misinformation given their inherent capabilities?** To the best of our knowledge, we are the first to explore such applications based on LVLM. We discover that LVLM can generally demonstrate satisfactory performance with its strong reasoning capability and proficiency in processing visual and textual information. However, challenges arise when external knowledge is necessary for an accurate prediction. In such cases, even reasoning-enhanced approaches have limited effectiveness in assisting LVLMs to make accurate decisions.

Points to these limitations, in this work, we propose LEMMA: **L**VLM-**E**nhanced **M**ulimodal **M**isinformation Detection with External Knowledge **A**ugmentation. The key motivation behind applying external knowledge is to provide evidence that can verify the authenticity of events, thereby enhancing the quality of LVLM's reasoning. Also, our approach maintains the advantages of both intuition and reasoning, assisting LVLM in crafting meticulous inferences based on the evidence from external knowledge while utilizing crucial cues unearthed through intuition to avoid excessive caution due to potential inaccuracies. Our experiments show that LEMMA significantly improves accuracy over the top baseline LVLM by **7%** and **13%** on the *Twitter* and *Fakeddit* datasets. In summary, the major contributions of this paper are as follows:

- We conduct a comprehensive empirical analysis of LVLM performance 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 to address the problem of multimodal misinformation detection.
- We design an ad-hoc external knowledge extraction module for LVLM to enhance the rigor and comprehensiveness of LVLM reasoning in multimodal misinformation detection tasks.

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 using image captioning model (Zhou et al., 2020), reflecting multimodality into the same latent space (Xue et al., 2021) 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. Moreover, (Hu et al., 2021) develops a GNN-based model to incorporate external knowledge for fake news detection. Given these insights, it becomes pertinent to explore the application of LVLM for this task. Their outstanding reasoning capabilities and profound real-world knowledge make them promising candidates for improving the accuracy of multimodal misinformation detection.

2.2 Knowledge-Augmented LLM/LVLM

Even though LLM holds a profound knowledge from vast pretrained corpus, they still present limited capability in some complicated tasks(Cao et al., 2023). To address such issues, (Guu et al., 2020; Lewis et al., 2020) pioneered retrieval-based methods to incorporate external resources such as Wikipedia into LLMs. In addition, (Wang et al., 2023) applies a knowledge retrieval module to improve LLM's performance in fact-checking tasks while (Baek et al., 2023) design a knowledge-augmented prompting method to help LLM in knowledge graph question answering task. Moreover, with the advent of LVLM, (Liu et al., 2023c) developed a multimodal assistant that acquires the ability to use external tools by being trained on multimodal instruction-following data. Since multimodal misinformation is usually detected by verifying real-world information, it is reasonable and promising to provide external knowledge to LVLM to achieve improved performance on such tasks.

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

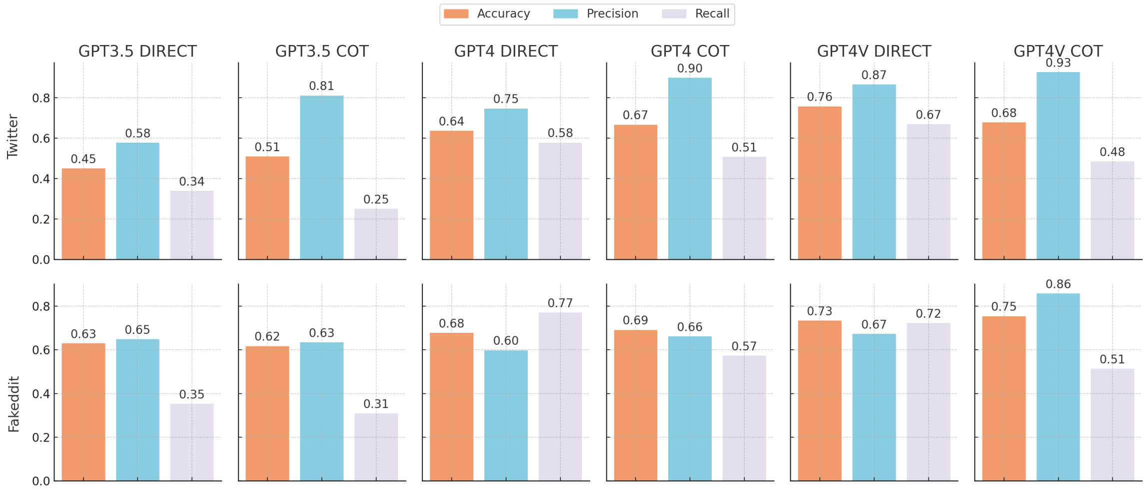

Figure 1: Comparison of performance metrics across different versions of GPT (GPT-3.5, GPT-4, and GPT-4V) and prompting methods (DIRECT and CoT) on two different datasets (*Twitter* and *Fakeddit*). It is observed that GPT-4V presents a superior performance compared to other LLMs.

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.

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 addition, to ensure a more comprehensive evaluation and to understand the evolution of LVLM, we also implement the Direct approach 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.

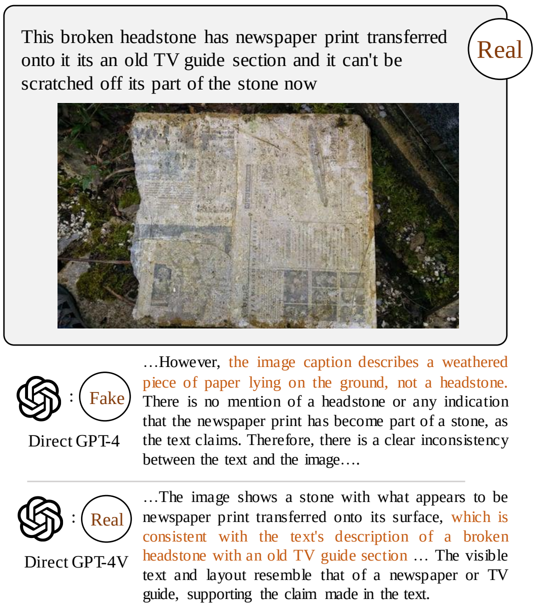

Figure 2: An example of a real *Fakeddit* post where GPT-4V makes a correct prediction based on successfully extracting cross-modal alignment, while GPT-4 fails.

3.2.4 Observation on Preliminary Result

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

1. **LVLM surpasses LLMs in comprehending cross-modal interaction**: Across both datasets and prompting methods, LVLM (GPT-4V) consistently demonstrates superior performance compared to LLMs (GPT-3.5 and GPT-4). This highlights LVLM notable capability of multimodality understanding. 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.

2. **In the absence of external evidence, reasoning-enhanced methods have very limited potential for performance improvement**: While CoT demonstrates superior precision on all versions of GPT, it simultaneously exhibits lower recall compared to the Direct method, which suggests a tendency towards over-conservatism. This conservative bias may stem from the inherent limitations of reasoning in the absence of substantial supporting evidence, underscoring a crucial trade-off between precision and recall in misinformation detection. For instance, 2 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.

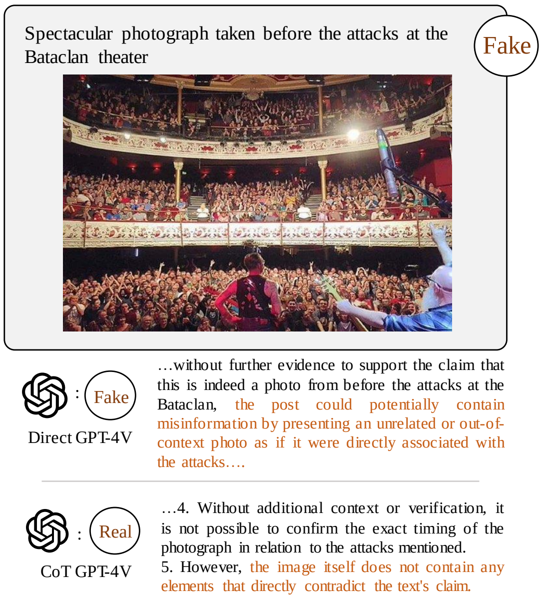

Figure 3: An example of a fabricated *Twitter* tweet that shares subtle discrepancies in two modalities, misleading GPT-4V to answer "presence of misinformation"

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. The LEMMA framework integrates multimodal inputs with external knowledge through a series of LVLM-based modules to enhance detection capabilities. The final predictions and reasoning of LEMMA are generated based on 1) the multimodal input, and 2) the filtered evidence extracted from external knowledge. We first delve into the initial stage inference in Section 4.1. Subsequently, we elucidate how we generate search

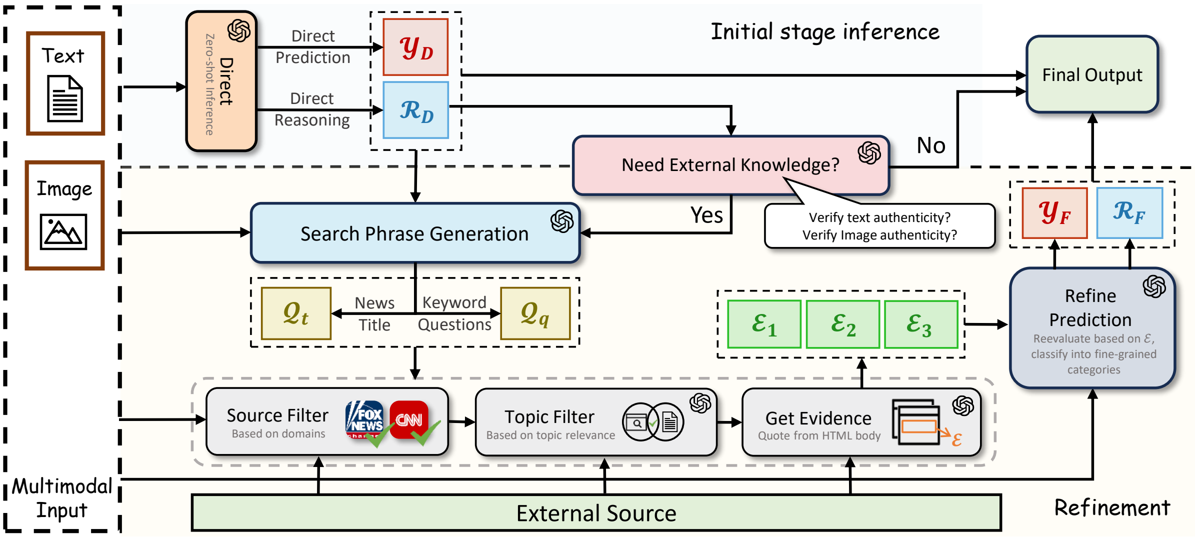

Figure 4: The pipeline of the proposed method (**LEMMA**). The process hinges on two key inputs: multimodal data and selectively filtered evidence gathered from external sources. Components marked with the OpenAI LOGO are developed using the LVLM (GPT-4V).

phrases to retrieve relevant 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.

4.1 Initial Stage Inference

In the initial phase, LVLM assesses whether news 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 in order to avoid an overly conservative judgment.

4.2 Search Phrase Generation

Recognizing the potential for conservative outputs in the absence of substantial evidence, LVLM procures external information to bolster its logical deductions. To maximize the relevance between the image-text pair $(\mathcal{I}, \mathcal{T})$ and the extracted evidence, we propose a twofold process: In the first step, LVLM is required to generate search phrases likely to yield pertinent results, empowering it to determine the external evidence required to refine its reasoning logic. Specifically, we supply LVLM with the image-text pair $(\mathcal{I}, \mathcal{T})$, the prediction $\mathcal{Y}_\mathrm{D}$ and the reasoning $\mathcal{R}_\mathrm{D}$ generated from initial stage inference. We then task LVLM with generating a title $\mathcal{Q}_\mathrm{t}$ for the news article and some search phrases $\mathcal{Q}_\mathrm{q}$, which are related to the content that needs to be verified. We construct the final searching queries set as $(\mathcal{Q}_\mathrm{t}, \mathcal{Q}_\mathrm{q})$. During the generation, we regularize LVLM with additional rules such that the generated queries would **1)** Be concise: The generated search phrase should consist of several keywords rather than forming a complete question. **2)** Be in English: Despite the possibility of textual inputs being in various languages, we mandate LVLM to consistently produce English titles and search phrases, as we observed that English queries result in better performance compared to using the original language of the textual input. Additionally, we append a "fake news" prefix to the queries to enhance the likelihood of articles refuting the under-tested multimodal input being returned.

4.3 External Knowledge Retrieval

In the second step, each query from the searching queries set $(\mathcal{Q}_\mathrm{t}, \mathcal{Q}_\mathrm{q})$ is fed into the DuckDuckGo Search API (DuckDuckGo, 2023) for external knowledge retrieval. The knowledge retrieval process comprises primarily two parts: **1)** Resource Distillation, which filters out untrusted websites and off-topic search results, and **2)** Evidence Extraction, which extracts relevant evidence from the filtered HTML body.

4.3.1 Resource Distillation

The resource distillation process unfolds in two main rounds: **1)** Source Filter: In the first round, the search engine (DuckDuckGo Search API) initially returns a set of sources $\mathcal{S}$ based on the top k relevance to the query set ($\mathcal{Q}_{\mathrm{t}}$, $\mathcal{Q}_{\mathrm{q}}$). Subsequently, a set of pre-collected domains of untrusted resources is additionally provided for the filtering, resulting in a refined set $\mathcal{S}'$. **2)** Topic Filter: Following source filtering, a second round of filtration based on topic relevance is applied to the remaining sources in $\mathcal{S}'$. Each source $\mathcal{S}_{\mathrm{i}} \in \mathcal{S}'$ comprises a web title and a brief description related to its context. The original news post text is then utilized with query context ($\mathcal{Q}_{\mathrm{t}}$, $\mathcal{Q}_{\mathrm{q}}$) to assist LVLM in assessing whether $\mathcal{S}_{\mathrm{i}}$ presents the related information appearing in the news post. Eventually, a further refined set $\mathcal{S}''$ is returned, containing the sources highly relevant and consistent with the information provided in the news post.

4.3.2 Evidence Extraction

Acknowledging that web pages may contain both irrelevant information and key evidence that either supports or refutes the multimodal input, our objective is to extract pertinent evidence from the HTML bodies of the filtered sources $\mathcal{S}''$. For each $\mathcal{S}_{\mathrm{i}} \in \mathcal{S}''$, we employ the Python module `newspaper3k` to extract the main content along with the publication date. By appending the title of the source $\mathcal{S}_{\mathrm{i}}$, we compose the full content of source $\mathcal{S}_{\mathrm{i}}$ as a triplet, comprising the title, publication date, and web content. Subsequently, we task LVLM with extracting key evidence from the full content of each $\mathcal{S}_{\mathrm{i}}$ that can either support or refute the original text $\mathcal{T}$. During the extraction, we regulate LVLM such that the evidence is directly quoted from the original HTML body and is as concise as possible while containing all information that potentially affects the authenticity of the text $\mathcal{T}$. Finally, we generate an evidence set $\mathcal{E}$ that consists of a list of extracted evidence.

4.4 Refined Prediction

With a set of extracted evidence $\mathcal{E}$ collected from external sources, it becomes possible to assess the factual accuracy of the raw text from news posts. Subsequently, the image-text pair ($\mathcal{I}$, $\mathcal{T}$) is reintroduced to the LVLM, accompanied with the evidence set $\mathcal{E}$. 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 news post into one of seven categories: 1) True, 2) Satire, 3) Misleading Content, 4) False Connection, 5) Imposter Content, 6) Manipulated Content, or 7) Unverified Content. Categories 2 through 7 correspond to different types of misinformation, while Category 1 indicates real news. If LVLM classifies the news post as Category 7, it will be asked to retain its inference from the initial stage, prioritizing conservatism over a potentially risky choice.

5 Experiments

5.1 Experiment Settings

For the evaluation of LEMMA, we establish a comparison with three categories of baseline models: **1) LLaVA:** We evaluate LLaVA-1.5-13B (Liu et al., 2023a), which is a state of art LVLM based on vision instruction tuning, by employing the Direct approach. **2) 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. **3) 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 5.9% in accuracy on *Twitter* and a notable 7% increase on *Fakeddit* when compared

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
	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
	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	**0.816**	**0.934**	0.741	0.825	**0.711**	0.924	**0.804**
	w/o *initial-stage infer*	0.718	0.866	0.598	0.707	0.621	0.877	0.727
	w/o *distillation*	0.808	0.880	**0.815**	**0.846**	0.706	0.801	0.749
Fakeddit	Direct (LLaVA)	0.663	0.588	**0.797**	0.677	0.777	0.558	0.649
	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
	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	**0.824**	0.835	0.727	**0.777**	**0.818**	0.895	**0.854**
	w/o *initial-stage infer*	0.803	0.857	0.692	0.766	0.786	0.891	0.830
	w/o *distillation*	0.795	**0.865**	0.654	0.758	0.748	0.914	0.829

Table 1: Performance comparison of baseline methods and LEMMA on *Twitter* and *Fakeddit* dataset. We show the result of five different baseline methods: Direct (LLaVA), Direct (GPT-4 with Image Summarization), CoT (GPT-4 with Image Summarization), Direct (GPT-4V), and CoT (GPT-4V). 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.

to the best-performing baseline. Moreover, our approach demonstrates superior capability in balancing precision and recall, reaching 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. In addition, LEMMA exhibits robust performance across different datasets, confirming its reliability and effectiveness in diverse contexts. This robustness is essential for practical applications, where a wide variety of data and scenarios need to be handled effectively.

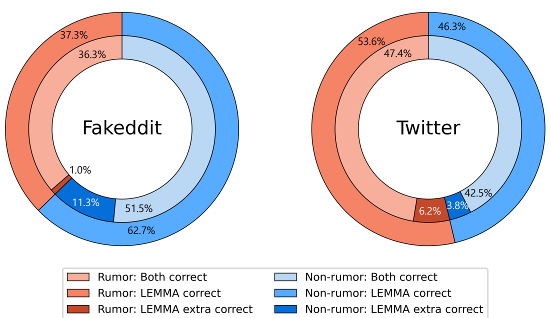

Figure 5: Comparison of the distribution of correct predictions between LEMMA and baseline (GPT-4V).

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 news posts. This led to a 9.8% lower accuracy on Twitter and a 2.4% 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*) *Distillation.* We also implement a version without resource distillation and evidence extraction, resulting in a 0.8% drop in accuracy on Twitter and a 2.9% drop on Fakeddit, suggesting that unprocessed external resources may contain confusing information that negatively affects LEMMA's judgment. However, omitting distillation improves recall and F1 on Twitter, We hypothesize that this is because the misinformation in the Twitter dataset does not rely on fine-grained evidence verification, thus the aggressiveness of the unfiltered version gains an advantage here.

5.4 Result Analysis

To delve deeper into the specific advantages of LEMMA, we conduct a statistical analysis to com-

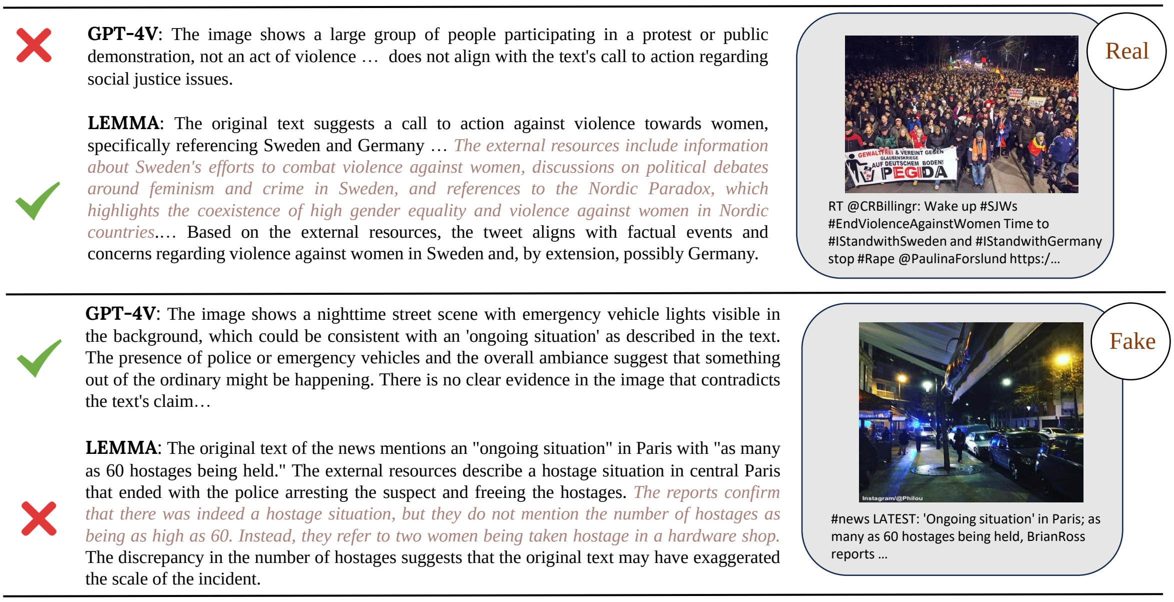

Figure 6: Two example posts in which LEMMA made an accurate prediction in the first, but erred in its forecast for the second

pare the accuracy distribution between LEMMA and Direct (GPT-4V). From Figure 5, 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 7% 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. In light of these statistical findings, the analysis of specific prediction examples in Figure 6 reveals the nuanced influence of external resources on LEMMA's predictive accuracy. From the first example, we observe that the external resources retrieved by LEMMA have provided crucial evidence for the prediction. However, in the second example, we observe that LVLM may be overly susceptible when the retrieved evidence does not conclusively validate the multimodal input.

6 Conclusion

In this study, we explored the inherent capability of LVLM in multimodal misinformation detection and discovered the significant importance of providing external information to enhance LVLM performance. Then we proposed a novel approach, LEMMA, which can effectively combine the intuitive and reasoning strengths of LVLM while addressing their factual grounding limitations. This exploration has revealed promising avenues for LVLM in the context of multimodal misinformation detection. Our experiments demonstrate that LEMMA significantly enhances accuracy compared to the top baseline LVLM, with improvements of **7%** and **13%** on the *Twitter* and *Fakeddit* datasets, respectively. While the scope remains for sophistication to the knowledge source interfaces and filtering, we believe LEMMA represents an extensible approach applicable to related interpretability-critical reasoning tasks at the intersection of vision, language, and verification. Future directions include expanding our approach to other multimodal formats (e.g. Text-Video pair) and developing more effective external sources filtering to further improve the quality of evidence.

7 Limitations

We recognize several limitations. 1) We did not explore other well-known LVLMs like Gemini and Kosmos-2 due to unavailable APIs. Different LVLMs may exhibit varying reasoning abilities in diverse cultural contexts, potentially impacting the generalizability of our proposed methods. 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 [1] 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.

Jinheon Baek, Alham Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In *Proceedings of the First Workshop on Matching From Unstructured and Structured Data (MATCHING 2023)*, pages 70–98, Toronto, ON, Canada. Association for Computational Linguistics.

Han Cao, Lingwei Wei, Mengyang Chen, Wei Zhou, and Songlin Hu. 2023. Are large language models good fact checkers: A preliminary study.

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

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.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

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.

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 (Volume 1: Long Papers)*, pages 754–763, Online. Association for Computational Linguistics.

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.

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.

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

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

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.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. 2023c. Llava-plus: Learning to use tools for creating multimodal agents.

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.

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.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2023. Factcheck-gpt: End-to-end fine-grained document-level fact-checking and correction of llm output. *ArXiv*, abs/2311.09000.

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.