XFacta: Contemporary, Real-World Dataset and Evaluation for Multimodal Misinformation Detection with Multimodal LLMs

Anonymous authors
Paper under double-blind review

Abstract

The rapid spread of multimodal misinformation on social media calls for more effective and robust detection methods. Recent advances leveraging multimodal large language models (MLLMs) have shown the potential in this challenge. However, it remains unclear exactly where the bottleneck of existing approaches lies (evidence retrieval *v.s.* reasoning), hindering the further advances in this field. On the dataset side, existing benchmarks either contain *outdated* events, leading to evaluation bias due to discrepancies with contemporary social media scenarios as MLLMs can simply memorize these events, or artificially *synthetic*, failing to reflect real-world misinformation patterns. Additionally, it lacks comprehensive analyses of MLLM-based model design strategies. To address these issues, we introduce XFacta, a contemporary, real-world dataset that is better suited for evaluating MLLM-based detectors. We systematically evaluate various MLLM-based misinformation detection strategies, assessing models across different architectures and scales, as well as benchmarking against existing detection methods. Building on these analyses, we further enable a semi-automatic detection-in-the-loop framework that continuously updates XFacta with new content to maintain its contemporary relevance. Our analysis provides valuable insights and practices for advancing the field of multimodal misinformation detection.

1 Introduction

A lie can travel halfway around the world before the truth can get its boots on—a statement that feels especially true in the age of social media. As platforms enable information to spread rapidly, humans face increasing challenges in identifying fake news online. Modern fake news is often multimodal, combining text with images that appear to support false or unrelated events, which makes detection more challenging. The rise of deepfake technology further lowers the barrier to creating such deceptive content. These developments highlight the need for more advanced and robust methods to automatically detect multimodal misinformation.

The emergence of multimodal large language models (MLLMs), with strong reasoning capabilities across both text and images, offers a promising direction for detecting multimodal misinformation. Recent studies have begun to explore this potential. Some methods (Qi et al., 2024; Liu et al., 2024a; Zeng et al., 2024; Shalabi et al., 2024) fine-tune a general-purpose MLLM on specific misinformation datasets to create task-specific models. Other approaches (Khaliq et al., 2024; Xuan et al., 2024; Liu et al., 2024b; Geng et al., 2024) adopt a zero-shot setting and rely on more powerful models such as GPT-4 or Gemini, which achieve better performance on existing misinformation datasets. In general, the existing MLLM-based misinformation detectors mimic human verification processes, which involves two main steps: *evidence retrieval*, where external information is retrieved from Internet to serve as evidence, then *reasoning*, where the news post and the retrieved evidence are systematically analyzed and combined to make final judgment.

Despite the promising results reported in these studies, it remains unclear exactly where the bottleneck of existing MLLM-based misinformation detection methods lies (evidence retrieval *v.s.* reasoning), hindering

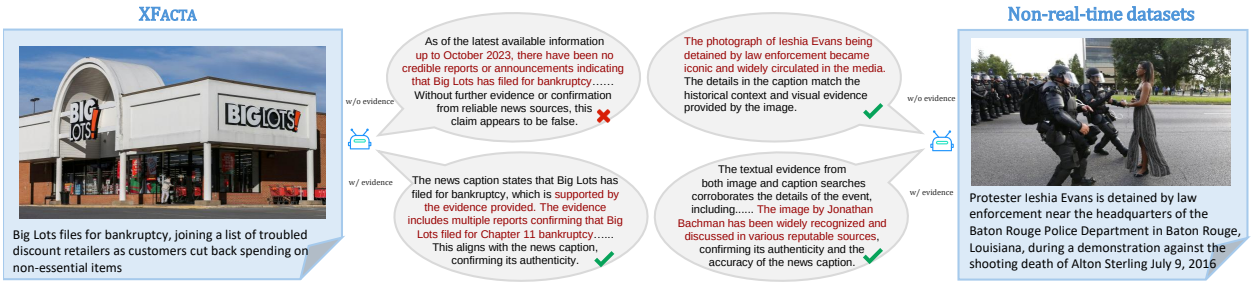

Figure 1: **Left**: An example from our dataset, where the MLLM (GPT-4o) must rely on evidence to judge real or fake. **Right**: An example from non-real-time datasets, where evidence matters less. Evaluating MLLM-based misinformation detectors on XFacta introduces less evaluation bias.

further advances in this field. From the **dataset perspective**, misinformation on real-world social media often involves novel and timely events that are absent from MLLMs' training data. Detecting these events requires models to actively retrieve evidence and reason thoroughly based on them. In contrast, existing misinformation benchmarks (Vlachos & Riedel, 2014; Wang, 2017; Thorne et al., 2018; Hanselowski et al., 2019; Khanam et al., 2021) contain mostly *outdated* data with events that may already exist in the training data of MLLMs, allowing models to rely simply on *memorization* rather than evidence-based reasoning. It introduces a significant evaluation bias as evidenced by an example shown in Fig. 1. In addition, some datasets (Luo et al., 2021; Chakraborty et al., 2023; Liu et al., 2024b; Shao et al., 2023; Aneja et al., 2021; Wu et al., 2026; Li et al., 2025b) are *synthetic*, meaning that misinformation samples are artificially constructed using AI models rather than collected from real-world sources. This limits their ability to reflect the complexity and strategies used by real misinformation creators. Regarding **technical approaches**, while existing studies typically focus on proposing new models or methods and demonstrating their effectiveness on specific datasets, it lacks *systematic analyses* and *rigorous comparisons* of different design choices for MLLM-based detection. Consequently, it still remains difficult today to identify best practices or generalizable insights for building reliable multimodal misinformation detectors.

In this paper, we address these limitations by curating a new misinformation dataset, named XFacta (collected from **X** (Twitter) and for **Fact**-checking). All data points are from after January 2024, ensuring its *contemporary* relevance (e.g., more recent than the October 2023 cutoff of GPT-4o). Moreover, they are sourced from rumor spreaders on social media, reflecting patterns observed in the *real world*. Based on this dataset, we conduct a systematic exploration of how to build an MLLM-based misinformation detector from the perspectives of evidence retrieval and reasoning, respectively. Additionally, we evaluate various MLLMs of different architectures and scales, as well as existing misinformation detection approaches. From these experiments and analyses, we provide valuable insights on MLLM-based misinformation detection. Building on these insights, we apply the resulting detector to flag new posts with preliminary assessments for human reviewers to verify and add to XFacta. This semi-automatic detection-in-the-loop cycle keeps the dataset up to date and prevents it from becoming outdated over time. We believe the XFacta dataset and our study results will serve as a useful benchmark for future research in multimodal misinformation detection.

To conclude, our contributions are:

- We curate a contemporary, real-world dataset for multimodal misinformation detection and integrate a semi-automatic detection-in-the-loop process to keep it continuously up to date, which will further advance MLLM-based detection research.
- With XFacta, we provide a comprehensive and in-depth analysis of developing a good MLLM-based misinformation detection model from two perspectives: evidence retrieval and reasoning, offering valuable insights to the field.
- We conduct a comprehensive evaluation of various MLLM-based misinformation detection strategies, assessing models across different architectures and scales, as well as benchmarking against existing detection methods.

Table 1: Comparison of different misinformation datasets. `Contemporary` refers to data published after January 1, 2024; `Real-world` means fake posts created by actual users, not artificially generated using AI models; and `Evidence-based annotations` mean there are annotations supported by sufficient evidence to verify the data.

Dataset	Multimodal	Contemporary	Real-world	Evidence-based annotations	Real Num	Fake Num
FEVER (Thorne et al., 2018)	✗	✗	✗	✓	93,367	43,107
LIAR (Wang, 2017)	✗	✗	✓	✓	7,085	5,751
LiveFact (Xu et al., 2026)	✗	✓	✓	✓	1,468	1,443
NewsCLIPpings (Luo et al., 2021)	✓	✗	✗	✗	816,922	816,922
Fakeddit (Nakamura et al., 2019)	✓	✗	✓	✗	527,049	628,501
Snopes+Reuters Zlatkova et al. (2019)	✓	✗	✓	✓	592	641
DGM[4] (Shao et al., 2023)	✓	✗	✗	✗	77,426	152,574
FACTIFY 3M (Chakraborty et al., 2023)	✓	✗	✗	✗	406,000	316,000
MMFakeBench (Liu et al., 2024b)	✓	✗	✗	✗	3,300	7,700
COSMOS (Aneja et al., 2021)	✓	✗	✗	✓	1,700	1,700
DeceptionDecoded (Wu et al., 2026)	✓	✗	✗	✓	4,000	8,000
Mocheg (Yao et al., 2023)	✓	✗	✓	✓	5,144	5,855
MediaEval (Boididou et al., 2016)	✓	✗	✓	✓	292	410
VERITE (Papadopoulos et al., 2023)	✓	✗	✓	✓	338	662
AVerImaTeC (Cao et al., 2025)	✓	✗	✓	✓	17	928
Post-4V (Geng et al., 2024)	✓	✓	✓	✓	81	105
XFacta (Ours)	✓	✓	✓	✓	1,300	1,300

2 Related Work

Datasets: Previous studies have introduced various unimodal text-based misinformation datasets (Vlachos & Riedel, 2014; Wang, 2017; Thorne et al., 2018; Hanselowski et al., 2019; Khanam et al., 2021). The rise of social media has led to increasing attention on multimodal misinformation detection, along with the release of various datasets (Nakamura et al., 2019; Zlatkova et al., 2019; Yao et al., 2023; Boididou et al., 2016; Papadopoulos et al., 2023). However, real-world misinformation datasets are typically either small in size or suffer from noisy annotations. Therefore, some other works (Luo et al., 2021; Chakraborty et al., 2023; Liu et al., 2024b; Shao et al., 2023; Aneja et al., 2021; Wu et al., 2026; Li et al., 2025b) leverage heuristic rules or AI models to synthesize datasets for misinformation detection. As a consequence of such synthesis, these datasets fail to capture real-world misinformation creators' complex patterns and strategies. In addition, all of the above datasets are not contemporary and often overlap with the training data of MLLMs, which prevents a fair and robust evaluation of MLLM-based misinformation detectors. Recent benchmarks have begun to consider the temporal nature of misinformation. LiveFact (Xu et al., 2026) emphasizes time-aware evaluation, but it focuses on text-only fake news detection. AVerImaTeC (Cao et al., 2025) targets multimodal verification in realistic settings, yet only a small portion of its data comes from post-2024 events. Post-4V (Geng et al., 2024) addresses this by collecting more recent examples, but its size is very limited, and data collection and processing details are underdocumented, making it less suitable as a widely accepted baseline. In contrast, our XFacta dataset ensures both contemporary and real-world characteristics, while maintaining a moderate scale that is sufficient to evaluate MLLMs in a zero-shot setting. In addition, our dataset provides detailed journalist evidence for fake news, which can help validate the reasoning paths of detection models. A multi-dimensional comparison across different datasets can be found in Table 1.

Models: Some traditional multimodal misinformation detectors Abdelnabi et al. (2022); Yuan et al. (2023); Brahma et al. (2023a); Aneja et al. (2023); Mu et al. (2023); Zhang et al. (2023); Brahma et al. (2023b); Yang et al. (2024); Cekinel et al. (2025) are trained and evaluated on specific datasets, such as the commonly used NewsCLIPpings dataset (Luo et al., 2021). With the emergence of open-source MLLMs, several recent works (Qi et al., 2024; Liu et al., 2024a; Zeng et al., 2024; Shalabi et al., 2024) have adopted a different approach by fine-tuning a pretrained MLLM on misinformation datasets, which achieves better performance. However, these methods often carry biases specific to their training data, which are not robust to new, more sophisticated misinformation emerging on social platforms. Therefore, several studies have explored more

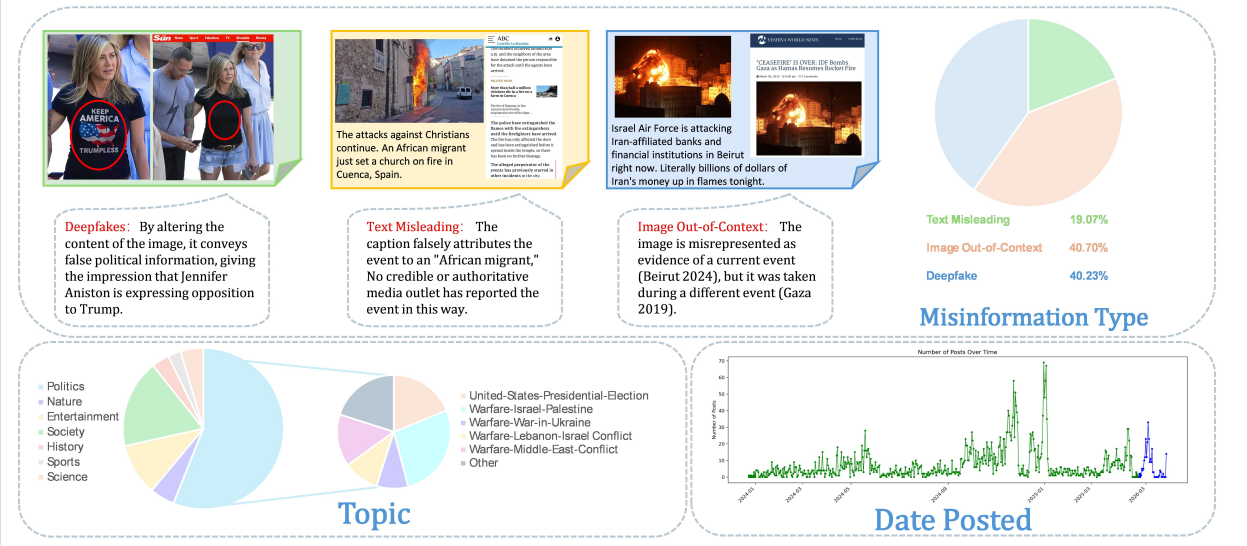

Figure 2: Examples and distribution of misinformation types, topics, and posting dates in XFacta.

powerful closed-source MLLMs and have achieved better results. However, these models are either claimed evidence-free (Geng et al., 2024; Wu et al., 2025; Wang et al., 2025), or evaluated on less updated or synthetic datasets (Khaliq et al., 2024; Xuan et al., 2024; Liu et al., 2024b; Jin et al., 2024; Li et al., 2025a; Tariq & Kementchedjhieva, 2026; Duwal et al., 2025; Shopnil et al., 2025), raising concerns about their effectiveness when deployed on evolving social media.

3 Our XFacta Dataset

Multimodal Misinformation Detection refers to assessing the authenticity of a news post that includes both supporting images and text. Formally, given a set of supporting images $I = \{I_1, \ldots, I_n\}$ and a text claim T, this task is to determine whether the post $\mathcal{P} = (I, T)$ is real or fake.

The supporting images I can make the text claim T seem more believable, even if they are unrelated or misleading, which makes detection much harder than in the unimodal setting. Therefore, most methods incorporate retrieved evidence $\mathcal{E} = (E_i, E_t)$ into their detection pipeline, where E_i and E_t are image-type and text-type evidence, respectively.

3.1 Data Source & Collection

Our dataset is sourced from X/Twitter. The real news posts are collected from authoritative news organizations including CNN, Fox News, The Guardian, and BBC. The fake news posts are curated from content flagged as false by BBC-certified journalists and X Community Notes.

We first collect fake news posts, as they are rarer and require careful identification, after which we gather about five times more real posts to ensure a diverse selection. This allows us to sample a subset of real posts that matches the fake posts in both quantity and distribution, reducing potential evaluation bias. We guarantee the distribution alignment in two aspects: (1) `Topic-aligned selection`, where we label the topic for both real and fake posts. We then ensure that the number of real and false posts per topic is the same, which helps reduce semantic differences by keeping the content semantically aligned. A detailed description of the topic of posts will be provided in Section 3.2. (2) `Image similarity selection`: While topic-aligned selection mainly matches the textual claims T, it does not explicitly control the image distribution I. To address this, we extract image embeddings using SigLIP (Zhai et al., 2023) and compute distances between fake posts and candidate real posts in the embedding space. We then use Optimal Transport to identify a subset of real posts whose image-feature distribution is best aligned with that of the fake posts. This design

aims to improve feature-level alignment between real and fake images and reduce image-only discriminability, thereby mitigating visual bias.

In addition, to ensure the reliability of news post labels, beyond the post content $\mathcal{P}$, each is provided with its metadata, including post URL, author id, date, topic, etc. For fake posts, we also collect flagging posts that give reasons and evidence for labeling as fake, while based on flagging posts, we also annotate the misinformation types, with more details provided in Section 3.2. We manually review each entry, and only those with clear evidence of misinformation are included in the dataset.

3.2 Data Statistics & Analysis

Our XFacta dataset contains a total of 2600 data points, including 1300 real posts and 1300 fake posts. For the convenience of model development, we randomly selected 120 real and 120 fake posts as the `Dev` set, while the remaining 2360 posts were used as the `Test` set. Notably, the bottom-right panel of Fig. 2 shows that most data were collected in 2025, whereas the blue-highlighted portion corresponds to additional samples collected in 2026 through the detector-in-the-loop expansion process described in Section 6. This contemporary setting allows XFacta to better reflect newly emerging misinformation trends beyond static or outdated benchmarks.

To better understand the dataset, we annotate each news post based on its topic and the types of misinformation in fake news. For topic classification, each post $\mathcal{P}$ is categorized into one of the following: *politics, society, entertainment, science, history, nature*, and *sports*, as shown in the bottom left corner of Fig. 2. Notably, political and conflict-related misinformation dominates but is also accompanied by other domains, which aligns closely with current global trends.

For fake posts, as illustrated at the top of Fig. 2, we assign one or more labels from three predefined misinformation types, based solely on explicit evidence provided in the collected flagging posts. We do not assign labels based on inference or assumptions beyond the provided evidence. The three error types are defined as follows:

- **Deepfakes**: The image is generated or digitally manipulated as identified by the flagging post.
- **Image Out-of-Context (OOC)**: The image is authentic but, according to the flagging post, originates from a different event than the one described in the accompanying text. This does not indicate whether the text is true or false.
- **Text Misleading**: The textual content conveys a claim that has been explicitly identified as false by the flagging post. This does not indicate whether the image is authentic or relevant.

By annotating each fake post with these finer-grained misinformation labels, we achieve a more nuanced understanding of the characteristics of multimodal misinformation, and enable a more detailed analysis of a misinformation detector's performance across different misinformation types.

4 How to Build a Good MLLM-based Misinformation Detector?

In this section, we explore different design strategies for MLLM-based misinformation detection on the XFacta dataset. We mainly investigate two questions: (1) How different types of evidence contribute to misinformation detection, and how we can better leverage them; (2) How different LLM reasoning approaches affect the model's prediction.

4.1 Analysis of Evidence Retrieval

4.1.1 Experiment Setup

For a given post $\mathcal{P}$ to be verified, we assume the retrieved evidence can assist the detection model in two main aspects: (1) verifying the authenticity of the event described in the post, and (2) verifying whether the accompanying image is used in an out-of-context manner. Based on these assumptions, we introduce eight evidence retrieval strategies designed to support these goals:

Table 2: Comparison of MLLM Performance with varying evidence retrieval approaches.

Evidence Type	GPT-4o			Gemini-2.0-flash			Qwen-vl-7b		
	Acc.	R. Acc.	F. Acc.	Acc.	R. Acc.	F. Acc.	Acc.	R. Acc.	F. Acc.
no evidence	70.8	50.8	90.8	71.7	78.3	65	60.8	76.7	44.4
Google Search									
① $T \rightarrow E_t$	87.1	97.5	76.7	81.3	98.3	64.2	59.7	82.7	38.1
② $T \rightarrow E_i$	81.7	75.8	87.5	77.9	90.8	65	62.2	84.2	40.2
③ $I \rightarrow E_t$	77.9	70	85.8	78.8	83.3	74.1	55.7	71.7	39.3
④ Query $\rightarrow E_i$	69.2	51.7	86.7	71.9	76.5	67.2	55.8	91.2	20
⑤ Query $\rightarrow E_t$	77.5	80	75	77.7	83.9	71.7	56.1	63.9	48.1
DuckDuckGo Search									
⑥ $T \rightarrow E_t$	84.2	94.2	74.2	79.2	97.5	60.8	64	91	37
⑦ $T \rightarrow E_i$	76.3	64.2	88.3	76.7	87.5	65.8	53	79	26.5
⑧ $T \rightarrow E_{news}$	84.2	80	88.3	75.3	92.4	58.3	68.4	84.4	52.9

Table 3: Comparison of different evidence post-processing methods with GPT-4o.

Method	Acc.	R. Acc.	F. Acc.
$T \rightarrow E_t$	87.1	97.5	76.5
Domain Filter	88.3	98.3	78.3
Evidence Extraction	87.5	95.8	79.2
$T \rightarrow E_i$	81.7	75.8	87.5
Domain Filter	83.8	80.8	86.7
Evidence Extraction	82	77	87
$I \rightarrow E_t$	77.9	70	85.8
Domain Filter	79.6	72.5	86.7
Evidence Extraction	81.3	75	87.5

- ① **Unimodal Evidence**: Using the post text T to retrieve textual evidence E_t to support Aspect (1). It mimics how humans verify news by searching for relevant information online.
- ②-③ **Cross-modal Evidence**: Using the post text T and image I separately to retrieve image-type evidence E_i (strategy ②) and text-type evidence E_t (stragety ③), following the cross-modal retrieval approach in Abdelnabi et al. (2022) to support Aspect (2).
- ④-⑤ **LLM Querying**: Using an LLM to generate questions about uncertain or suspicious details in the post, then forming search queries to retrieve image-type evidence E_i (strategy ④) and text-type evidence E_t (strategy ⑤). This simulates how humans investigate unclear claims by asking targeted questions.
- ⑥-⑧ **DuckDuckGo Variants**: To explore how different search engines influence retrieval results, we replace the search engine used strategies ⑥ and ⑦ with DuckDuckGo in strategies ① and ②, respectively, We also use DuckDuckGo's "search news" for news evidence E_{news} (strategy ⑧), investigating whether it can retrieve more authoritative evidence.

Additionally, we believe that post-processing can help clean the evidence to reduce its noise. Here, we propose two methods for evidence post-processing inspired by Xuan et al. (2024):

- **Domain Filter**: Filtering out evidence from untrustworthy domains.[1]
- **Evidence Extraction**: Using an MLLM (GPT-4o in our paper) to select parts of the evidence that are highly relevant to the news post and remove irrelevant parts.

To evaluate the impact of each evidence type, we first run the model without evidence, relying only on an MLLM's internal knowledge. Then, we add each of the eight evidence types separately and compare the results against the no-evidence baseline and with each other. We use Chain-of-Thought (CoT) (Wei et al., 2022) prompting to obtain interpretable reasoning outputs instead of simple binary decisions. Experiments are conducted on the `Dev` set using three models with different scales: GPT-4o (Yang et al., 2023), Gemini-2.0-Flash (Team et al., 2023), and Qwen-VL-7B (Wang et al., 2024), to reduce model-specific bias. For post-processing, we test the performance with and without each strategy on GPT-4o. We report three metrics: overall accuracy (Acc.), accuracy on real posts (R. Acc.), and accuracy on fake posts (F. Acc.). The model also outputs a confidence score (0–100), and we report average confidence (Avg. Conf.) in certain tables to reflect prediction certainty.

4.1.2 Results and Analysis

Table 2 and 3 present the performance of different evidence retrieval and post-processing strategies, respectively. Table 4 presents a more detailed comparison across misinformation types for fake posts. We summarize several key observations as follows.

[1]Evidence from domains used in dataset curation is excluded by default to avoid leakage.

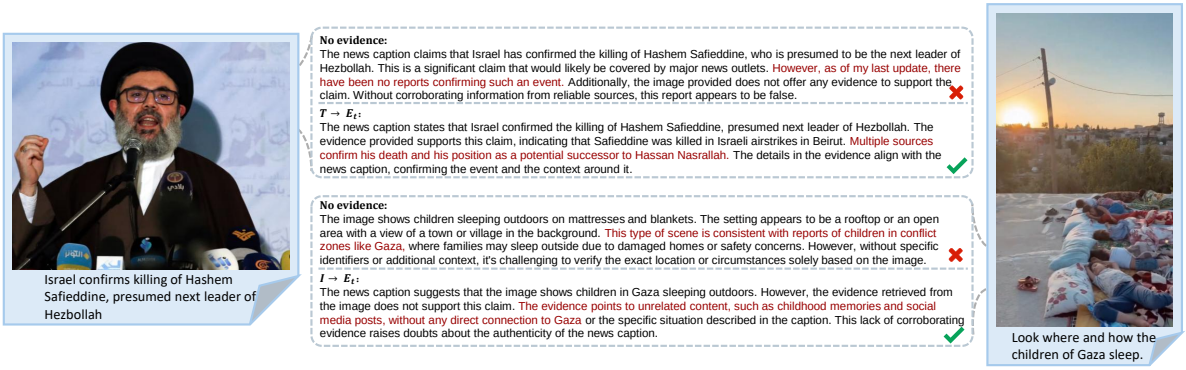

Figure 3: Examples of how $T \rightarrow E_t$ and $I \rightarrow E_t$ correct the no-evidence detection error.

1. All types of evidence consistently improves accuracy over the no-evidence baseline. Without evidence, models exhibit notable differences in their behavior: GPT is more conservative, whereas Gemini and Qwen are more inclined to label posts as real. With evidence, their classifications become more balanced, showing the importance of external evidence in misinformation detection.

2. $T \rightarrow E_t$ (strategy ①) substantially boosts performance, especially for real posts. This is expected—even for humans, real news is more likely to be supported by online evidence and thus easier to verify. See left of Fig. 3 for example. However, accuracy for fake posts does not notably improve, and even declines slightly for GPT-4o. We attribute this to OOC misinformation, where $T \rightarrow E_t$ provides no information about the image I, and the strong support for T in E_t misleads the model to flip its originally correct prediction of fake.

3. $I \rightarrow E_t$ (strategy ③) is more effective than $T \rightarrow E_i$ (strategy ②) for out-of-context misinformation. Although $T \rightarrow E_i$ shows higher overall accuracy in Table 2, manual inspection reveals that $I \rightarrow E_t$ better detects out-of-context cases. This is because $I \rightarrow E_t$ retrieves webpages directly containing the query image and extracts highly relevant text, while $T \rightarrow E_i$ conducts a fuzzy search based on the caption and often retrieves loosely related images. In addition, textual evidence is also more informative in these cases, since image-based comparisons are often limited to coarse features like general scenes or people. These superficial similarities are usually preserved in out-of-context misinformation, making it hard to detect manipulation through image-type evidence. Point 5 further analyzes strategy ③ on fake posts.

4. LLM-generated queries (strategies ④ and ⑤) are less effective than direct caption searches. In most cases, an LLM is not able to generate highly targeted queries; most of them are simply paraphrases of the original caption. Searching with such paraphrased versions is thus less accurate than directly using the caption T itself to retrieve evidence. In certain cases of fake posts, if the questions or doubts raised by the LLM fail to target the actual reason why the post is fake, the retrieved evidence can even lead the model to confidently make an incorrect judgment, as further analyzed in the next point.

5. Across different misinformation types, $I \rightarrow E_t$ (strategy ③) provides consistently informative evidence especially for identifying fake posts. Unlike earlier analyses based on the overall accuracy, analyzing fine-grained misinformation types for fake posts demands careful consideration beyond accuracy alone. GPT-4o tends to conservatively classify ambiguous posts as fake even without additional evidence, and this can inflate the accuracy of fake posts. Hence, average confidence scores become essential because they indicate whether the retrieved evidence provides clear and informative knowledge that truly helps the model's judgment. As shown in Table 4, strategy ③ not only achieves high accuracy but also consistently maintains high confidence across Deepfakes and Image OOC categories. Although Query $\rightarrow E_t$ (strategy ⑤) shows slightly superior combined performance in the Text Misleading category, it causes the model to make highly confident but incorrect predictions in Image OOC cases, significantly reducing its overall utility. Therefore, $I \rightarrow E_t$ remains the optimal evidence retrieval strategy across various misinformation types.

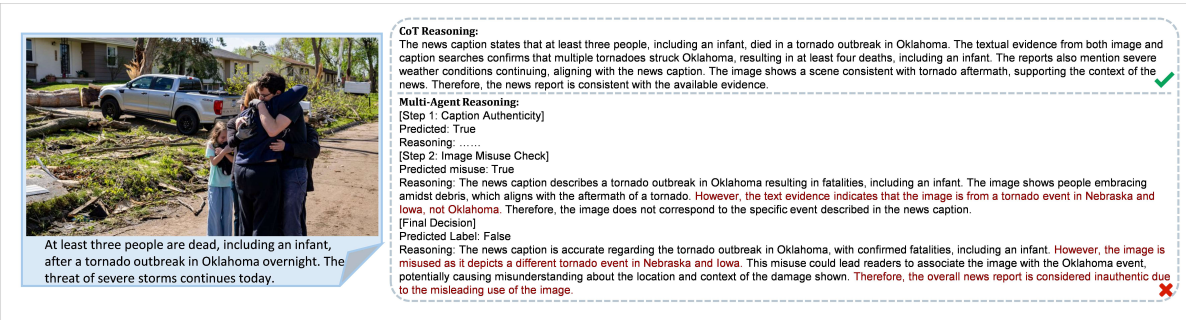

Figure 4: Multi-Agent reasoning can detect image "misuse" in the news post from CNN. We believe this "overly strict" behavior is actually beneficial for reliable misinformation detection.

Table 4: Comparison of each evidence retrieval strategies across misinformation types with GPT-4o.

Evidence Type	**Deepfakes**		**Image OOC**		**Text Misleading**	
	Acc.	Avg. Conf.	Acc.	Avg. Conf.	Acc.	Avg. Conf.
no evidence	89.7	87.4	93.8	77	91.1	82.6
Google Search						
① $T \rightarrow E_t$	79.3	87.8	77.1	84	80.4	87
② $T \rightarrow E_i$	93.1	85.9	85.4	81	87.5	83
③ $I \rightarrow E_t$	100	88.5	83.3	85.2	80.4	88.2
④ Query $\rightarrow E_i$	89.7	88.8	79.2	82.9	89.3	85.4
⑤ Query $\rightarrow E_t$	86.2	91.2	60.4	88.7	82.1	89.6
DuckDuckGo Search						
⑥ $T \rightarrow E_t$	79.3	90	64.6	85.5	82.1	86.1
⑦ $T \rightarrow E_i$	89.7	85.9	83.3	79.6	91.1	80.9
⑧ $T \rightarrow E_{news}$	93.1	84.8	81.3	80.6	92.9	80.8

6. DuckDuckGo provides lower-quality evidence than Google (strategis ⑥-⑧). Experiments indicate that evidence retrieved using DuckDuckGo consistently yields lower performance compared to Google Search. Additionally, $T \rightarrow E_{news}$ did not produce the expected improvements.

7. Domain Filter can mitigate evidence noise. As shown in Table 3, domain filter can improve accuracy in general, suggesting that evidence from low-credibility websites is indeed noisy and potentially misleading.

8. LLM-Based Evidence Extraction can mitigate evidence noise. We inspected the extraction results and found that the LLM can successfully retain the key information needed to detect misinformation and filter out some irrelevant evidence, which leads to improved detection accuracy as shown in Table 3, especially for $I \rightarrow E_t$. However, it is important to note that evidence extraction introduces a huge token overhead.

4.2 Analysis of Reasoning

4.2.1 Experiment Setup

We use $T \rightarrow E_t$ (strategy ①) and $I \rightarrow E_t$ (strategy ③) in the reasoning stage, as they can complement each other well. We also apply domain filter to reduce evidence noise, but skip evidence extraction to better assess the reasoning ability on noisy evidence pieces. We test four reasoning strategies, including **CoT** (Wei et al., 2022), **Prompt Ensembles** (Geng et al., 2024), **Self Consistency** (Wang et al., 2022), and **Multi-Agent Reasoning**. See the appendix for additional details.

Table 5: Comparison of MLLM Performance with various reasoning methods on the Dev set.

Reasoning Method	GPT-4o			Gemini-2.0-flash			Qwen-vl-7b		
	Acc.	R. Acc.	F. Acc.	Acc.	R. Acc.	F. Acc.	Acc.	R. Acc.	F. Acc.
Chain of Thought	88.3	98.3	78.3	83.8	98.3	69.2	54.8	84.2	24.1
Prompt Ensembles	90	100	80	85.4	98.3	72.5	67.1	90	44
Self Consistency	88.3	97.5	79.2	86.7	98.3	75	61	64	58.2
Multi-Agent Reasoning	91.3	91.7	90.8	81.3	90	72.5	62.1	78.4	45.9

Table 6: Comparison of detection performance across different LLMs on the Test set.

Model	Scale	Acc.	R. Acc.	F. Acc.
GPT	GPT-4o-mini	83	84.6	81.3
	GPT-4o	88.6	87.6	89.6
Gemini	Gemini-2.0-lite	76.2	77.2	75.2
	Gemini-2.0-flash	78.9	83.6	74.2
Qwen	Qwen-vl-7b	65	80.9	48.5
	Qweb-vl-72b	81	82.3	79.6

Table 7: Comparison of different misinformation detection methods on the Test set. *: 284 samples are excluded due to the ambiguous output.

Methods	Training Set / LLM	Acc.	R. Acc.	F. Acc.
SENs	NewsCLIPpings	52.6	58.4	46.7
Mocheg	Mocheg	51.7	57.6	45.5
HAMMER	DGM[4]	57.2	78.2	36.1
Sniffer	NewsCLIPpings	56.1	71	41.1
MMFakeBench*	GPT-4o	68.2	61.5	75.6
LEMMA	GPT-4o	77.3	63.9	90.8
Ours	GPT-4o	88.8	87.2	90.4

4.2.2 Results and Analysis

We summarize several key observations below according to the results reported in Table 5.

1. The stronger the MLLM, the less it is affected by different reasoning methods. Stronger models like GPT-4o usually have good reasoning ability by default and have similar accuracy across different reasoning techniques.

2. Different model architectures show different preferences for different reasoning methods. Therefore, in practice, deploying an MLLM-based misinformation detector should involve testing various reasoning methods, especially for smaller MLLMs, to achieve better performance.

3. For GPT-4o, multi-agent reasoning has the overall best balanced accuracy. For the best performing model GPT-4o, by manually inspecting the reasoning paths across various strategies, we find that multi-agent reasoning consistently provides the clearest and most structured reasoning. Particularly, its accuracy in detecting fake posts is superior to other methods. However, its accuracy on real posts is not that good. Interestingly, we found that some real posts from reputable news sources may use images from unrelated events (which we do not consider as misinformation because there is no intention to mislead). Multi-Agent reasoning can identify and flag these cases as fake due to image mismatch. We believe this "overly strict" behavior is actually beneficial for reliable misinformation detection. An example can be found in Fig. 4.

5 Further Evaluations on XFacta

Comparison of Different MLLMs. We evaluate the performance of various MLLMs on our Test set. Specifically, we analyze performance differences across closed-source models (GPT and Gemini), as well as open-source models (Qwen), both across different model scales. We use the evidence types following Section 4.2 and use the multi-Agent reasoning strategy. Results are shown in Table 6. For the same model architecture, larger models always achieve higher accuracy.

Comparison of Existing Multimodal Misinformation Detection Methods. We perform a horizontal comparison with existing multimodal misinformation approaches using our Test set, including models trained from scratch: SENs (Yuan et al., 2023), Mocheg (Yao et al., 2023), and HAMMER (Shao et al., 2023); methods fine-tuning MLLMs: Sniffer (Qi et al., 2024), and zero-shot methods using closed-source MLLMs: MMFakeBench (Liu et al., 2024b), LEMMA (Xuan et al., 2024). Results in Table 7 show that specialist

Table 8: Prompt sensitivity study.

Prompt Template	Direct	With Evidence
Ours	70.8	88.3
LEMMA-style	78.8	90.4
MMFakeBench-style	75.0	86.5

Table 9: Temporal performance comparison.

Evidence Setting	Jan–Jun 2024			Jul 2024+		
	Acc.	R.Acc	F.Acc	Acc.	R.Acc	F.Acc
w/o evidence	81.8	81.8	81.7	73.3	67.1	82.4
w/ evidence	83.6	84.8	82.4	84.05	83.59	84.73

methods that trained on a specific dataset suffer from severe generalization issues. In contrast, models that use GPT-4o demonstrate relatively good performance. Based the systematic analysis of evidence retrieval and reasoning strategies, our method outperforms these models, establishing SOTA accuracy on XFacta.

Comparison of Similar Datasets. We examine the dependency on evidence across similar datasets. To achieve this, we select a subset from (Papadopoulos et al., 2023), Snopes+Reuters Zlatkova et al. (2019), and NewsCLIPpings (Luo et al., 2021) dataset, respectively. We use the evidence types following Section 4.2 and apply CoT reasoning with GPT-4o. We report the results in Table 10 in the Appendix. Notably, for our XFacta dataset, GPT-4o does not work well without any evidence, confirming the need for a contemporary, real-world benchmark.

Prompt Sensitivity Analysis. To examine how prompt design affects our experimental results, We conduct a prompt sensitivity study on GPT-4o using three prompt templates: (1) our original prompt, (2) a LEMMA-style prompt adapted from Xuan et al. (2024), and (3) an MMFakeBench-style prompt adapted from Liu et al. (2024b). For each template, we evaluated two settings:

- **Direct**: classification without evidence.
- **With evidence**: classification with structured evidence produced by our evidence pipeline.

Moreover, to make the reproduction as fair and reliable as possible, we followed a strict alignment strategy when adapting the compared prompts to our setting. Since our experimental setup is not exactly the same as the original settings, we preserved their core components as much as possible. Specifically, we kept the same basic prompt structure, adopted the same definition of misinformation, and followed their original style in describing the caption, image, and evidence. Meanwhile, we removed prompt components that were not compatible with our pipeline. For example, LEMMA asks the model to first determine whether external knowledge is needed. As this mechanism is not part of our setting, we excluded it from the reproduced prompt. In this way, our reproduction retains only the shared and comparable parts of each method, while avoiding unfair differences caused by pipeline mismatch.

As shown in Table 8, when no evidence is provided, GPT-4o is sensitive to prompt wording. However, once structured evidence is provided, the performance gap across different prompt designs becomes much smaller, indicating that prompt sensitivity is greatly reduced. This suggests that the improvements mainly come from our evidence-based framework rather than prompt engineering alone.

Effect of Contemporaneity. To better understand how contemporaneity influences model behavior, we conducted a cross-time experiment. Specifically, we selected GPT-4.1 as the base model because its knowledge cutoff is June 2024. Based on this cutoff, we divided the test set into two non-overlapping subsets: 493 samples from January–June 2024, which the model may have partially seen during training, and 1,907 samples from July 2024 onward, which the model has no access to. We evaluated both subsets using the same Chain-of-Thought reasoning strategy, and measured detection accuracy under two conditions: without retrieved evidence and with retrieved evidence. The results, shown in Table 9, reveal a clear temporal performance gap. In the "no evidence" setting, the model exhibits a substantial accuracy drop on the post-cutoff subset, indicating that its detection ability weakens when misinformation falls outside its knowledge period. When retrieved evidence is provided, however, this performance gap narrows significantly. This suggests that for contemporary misinformation, the model relies more on external evidence retrieval rather than internal memorized knowledge.

Detector Effectiveness on More Recent and Out-of-Distribution Data. We evaluate whether the misinformation detector trained on the original XFacta dataset remains effective when applied to more

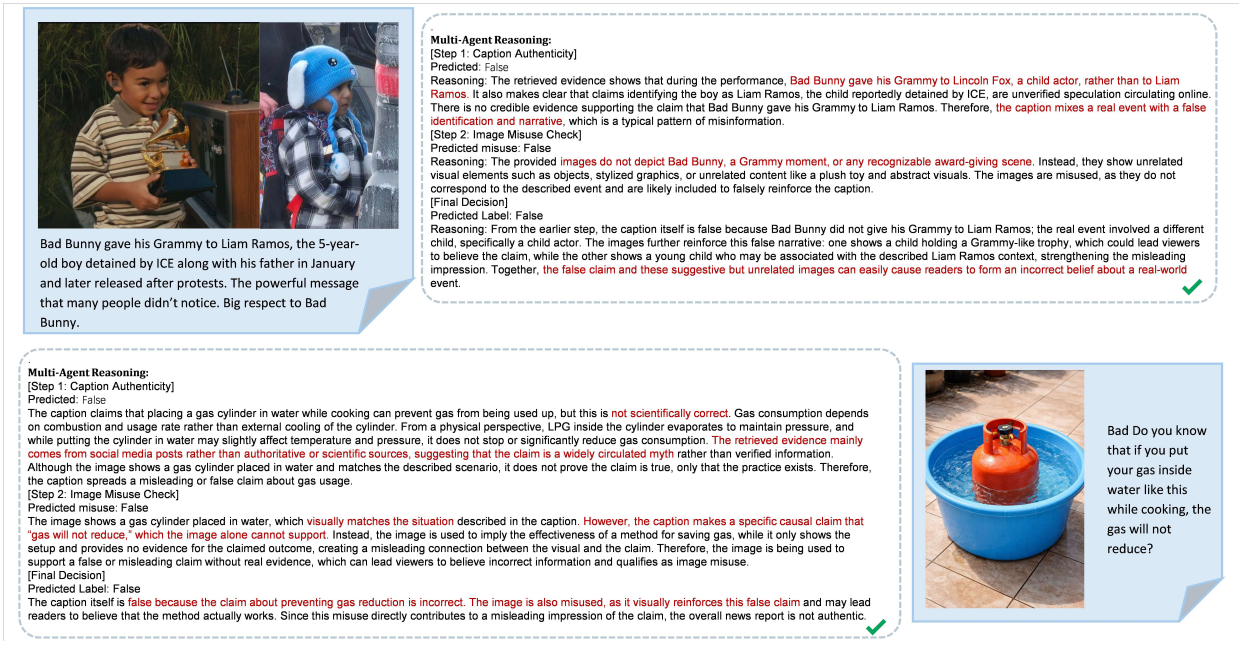

Figure 5: Examples Collected by the Detector-Assisted Expansion Process

recent and out-of-distribution social media content. First, we choose Snopes as the testbed since, compared to X, it is out-of-distribution. Moreover, the website provides real/fake annotations which are provided by professional journalists, thus can serve as a reference for evaluating the performance of the detector. We collect 1,200 fact-checked news items (600 real and 600 fake) from Snopes between July 2024 and July 2025, which are more recent than original XFACT dataset. Our resulting detector achieves an overall accuracy of 89.2% on this dataset (85.5% on true and 93.0% on false), showing that the model works well on both newest and out-of-distribution data.

6 Closing the Loop: Detector-Assisted Dataset Expansion

In this section, we demonstrate how our detector, which has been validated on the original XFACTA dataset, can be effectively used to support dataset expansion. Previous experiments show that the detector maintains stable performance on both more recent and out-of-distribution data, suggesting that it generalizes well to continuously emerging, previously unseen content. Therefore, it can be integrated into the dataset collection pipeline to assist human reviewers in verifying misinformation, enabling a detection-in-the-loop framework that accelerates and scales up the data curation process.

To verify this idea, we conduct a case study as a proof of concept. This time, we do not rely on journalist-flagged posts or posts from official news accounts for real/fake references. Instead, we select several accounts that regularly post about trending or controversial topics and that have a sizable follower base. We crawl and identify 500 posts between June 2025 and July 2025 from these accounts. Among them, 265 posts are identified by our detector as fake and 235 as real. For each prediction, the detector also generates an explanation to support its decision. To further assess the reliability of this detector-assisted collection process, we randomly select 200 posts from the collected candidates and manually verify their labels. The detector achieves an accuracy of 78% on posts predicted as real and 90% on posts predicted as fake, suggesting that the detector can provide a useful first-stage filtering signal, especially for identifying potential misinformation candidates. The explanations can further assist human reviewers in verifying the predictions more efficiently and deciding whether to incorporate the posts into XFACT. An example can be found in Figure. 5. The additional samples will be incorporated into the XFACTA dataset.

7 Conclusion

In this paper, we introduced XFacta, a contemporary, real-world dataset for multimodal misinformation detection. Using this dataset, we analyze how to build an effective MLLM-based misinformation detector from two perspectives: evidence retrieval and reasoning. Our experiments offer practical insights into developing robust detection systems. Furthermore, we implement a semi-automatic detection-in-the-loop cycle to continuously update XFacta with newly flagged content. We also benchmark SOTA MLLMs and existing detection methods in a more realistic setting using our dataset. We believe that XFacta and our findings will foster future research in multimodal misinformation detection.

Ethics Statement

Our research adheres to the guidelines set forth by the Twitter Developer Terms[2]. We ensure that our data collection and use comply with these terms, including the appropriate use of the Twitter API. While we plan to release our dataset for research purposes, we will do so in a manner that adheres to all applicable rules and guidelines.

Our study focuses on detecting multimodal misinformation, a significant issue in the digital age. By identifying and mitigating the spread of misinformation, our work contributes positively to the integrity of information on the web. Since our dataset consists of internet fake news posts, some posts may contain offensive content. However, the positive contributions of our research in reducing misinformation far outweigh the potential negatives.

References

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14940–14949, 2022.

Shivangi Aneja, Chris Bregler, and Matthias Nießner. Cosmos: Catching out-of-context misinformation with self-supervised learning, 2021. URL https://arxiv.org/abs/2101.06278.

Shivangi Aneja, Chris Bregler, and Matthias Nießner. Cosmos: catching out-of-context image misuse using self-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 14084–14092, 2023.

Christina Boididou, Symeon Papadopoulos, Duc Tien Dang Nguyen, G. Boato, Michael Riegler, Andreas Petlund, and Ioannis Kompatsiaris. Verifying multimedia use at mediaeval 2016. 10 2016.

Debarshi Brahma, Amartya Bhattacharya, Suraj Nagaje Mahadev, Anmol Asati, Vikas Verma, and Soma Biswas. Dpod: Domain-specific prompt tuning for multimodal fake news detection. *arXiv preprint arXiv:2311.16496*, 2023a.

Debarshi Brahma, Amartya Bhattacharya, Suraj Nagaje Mahadev, Anmol Asati, Vikas Verma, and Soma Biswas. Leveraging out-of-domain data for domain-specific prompt tuning in multi-modal fake news detection. *arXiv preprint arXiv:2311.16496*, 2023b.

Rui Cao, Zifeng Ding, Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. Averimatec: A dataset for automatic verification of image-text claims with evidence from the web, 2025. URL https://arxiv.org/abs/2505.17978.

Recep Firat Cekinel, Pinar Karagoz, and Çağrı Çöltekin. Multimodal fact-checking with vision language models: A probing classifier based solution with embedding strategies. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4622–4633, Abu Dhabi, UAE, January

[2] https://developer.x.com/en/developer-terms/more-on-restricted-use-cases

2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.310/.

Megha Chakraborty, Khusbu Pahwa, Anku Rani, Adarsh Mahor, Aditya Pakala, Arghya Sarkar, Harshit Dave, Ishan Paul, Janvita Reddy, Preethi Gurumurthy, et al. Factify3m: A benchmark for multimodal fact verification with explainability through 5w question-answering. *arXiv preprint arXiv:2306.05523*, 2023.

Sharad Duwal, Mir Nafis Sharear Shopnil, Abhishek Tyagi, and Adiba Mahbub Proma. Evidence-grounded multimodal misinformation detection with attention-based gnns, 2025. URL https://arxiv.org/abs/2505.18221.

Jiahui Geng, Yova Kementchedjhieva, Preslav Nakov, and Iryna Gurevych. Multimodal large language models to support real-world fact-checking, 2024. URL https://arxiv.org/abs/2403.03627.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. A richly annotated corpus for different tasks in automated fact-checking. *arXiv preprint arXiv:1911.01214*, 2019.

Yiqiao Jin, Minje Choi, Gaurav Verma, Jindong Wang, and Srijan Kumar. Mm-soc: Benchmarking multimodal large language models in social media platforms. *arXiv preprint arXiv:2402.14154*, 2024.

M Abdul Khaliq, Paul Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models. *arXiv preprint arXiv:2404.12065*, 2024.

Zeba Khanam, BN Alwasel, H Sirafi, and Mamoon Rashid. Fake news detection using machine learning approaches. In *IOP conference series: materials science and engineering*, volume 1099, pp. 012040. IOP Publishing, 2021.

Fanxiao Li, Jiaying Wu, Canyuan He, and Wei Zhou. CMIE: Combining MLLM insights with external evidence for explainable out-of-context misinformation detection. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9342–9354, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.487. URL https://aclanthology.org/2025.findings-acl.487/.

Haiyang Li, Yaxiong Wang, Shengeng Tang, Lianwei Wu, Lechao Cheng, and Zhun Zhong. Towards unified multimodal misinformation detection in social media: A benchmark dataset and baseline, 2025b. URL https://arxiv.org/abs/2509.25991.

Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng He. Fka-owl: Advancing multimodal fake news detection through knowledge-augmented lvlms. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10154–10163, 2024a.

Xuannan Liu, Zekun Li, Peipei Li, Shuhan Xia, Xing Cui, Linzhi Huang, Huaibo Huang, Weihong Deng, and Zhaofeng He. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms, 2024b. URL https://arxiv.org/abs/2406.08772.

Grace Luo, Trevor Darrell, and Anna Rohrbach. Newsclippings: Automatic generation of out-of-context multimodal media. *arXiv preprint arXiv:2104.05893*, 2021.

Michael Mu, Sreyasee Das Bhattacharjee, and Junsong Yuan. Self-supervised distilled learning for multimodal misinformation identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2819–2828, 2023.

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

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. Verite: A robust benchmark for multimodal misinformation detection accounting for unimodal bias, 2023. URL https://arxiv.org/abs/2304.14133.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13052–13062, 2024.

Fatma Shalabi, Hichem Felouat, Huy H Nguyen, and Isao Echizen. Leveraging chat-based large vision language models for multimodal out-of-context detection. In *International Conference on Advanced Information Networking and Applications*, pp. 86–98. Springer, 2024.

Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and grounding multi-modal media manipulation, 2023. URL https://arxiv.org/abs/2304.02556.

Mir Nafis Sharear Shopnil, Sharad Duwal, Abhishek Tyagi, and Adiba Mahbub Proma. Mirage: Agentic framework for multimodal misinformation detection with web-grounded reasoning, 2025. URL https://arxiv.org/abs/2510.17590.

Snopes. URL https://www.snopes.com/.

Amina Tariq and Yova Kementchedjhieva. REVEAL: Retrieval-enhanced verification for multimodal fact-checking. In Mubashara Akhtar, Rami Aly, Rui Cao, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos (eds.), *Proceedings of the Ninth Fact Extraction and VERification Workshop (FEVER)*, pp. 108–113, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-365-4. doi: 10.18653/v1/2026.fever-1.8. URL https://aclanthology.org/2026.fever-1.8/.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pp. 18–22, 2014.

Bing Wang, Ximing Li, Yanjun Wang, Changchun Li, Lin Yuanbo Wu, Buyu Wang, and Shengsheng Wang. Enhancing multimodal misinformation detection by replaying the whole story from image modality perspective, 2025. URL https://arxiv.org/abs/2511.06284.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 422–426, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2067. URL https://aclanthology.org/P17-2067/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Jiaying Wu, Fanxiao Li, Zihang Fu, Min-Yen Kan, and Bryan Hooi. Seeing through deception: Uncovering misleading creator intent in multimodal news with vision-language models, 2026. URL https://arxiv.org/abs/2505.15489.

Junjie Wu, Yumeng Fu, Chen Gong, and Guohong Fu. Insight-a: Attribution-aware for multimodal misinformation detection, 2025. URL https://arxiv.org/abs/2511.21705.

X Community Notes. URL https://communitynotes.x.com/guide/en/about/introduction.

X/Twitter. URL https://x.com/.

Cheng Xu, Changhong Jin, Yingjie Niu, Nan Yan, Yuke Mei, Shuhao Guan, Liming Chen, and M-Tahar Kechadi. Livefact: A dynamic, time-aware benchmark for llm-driven fake news detection, 2026. URL https://arxiv.org/abs/2604.04815.

Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R Fung, and Heng Ji. Lemma: Towards lvlm-enhanced multimodal misinformation detection with external knowledge augmentation. *arXiv preprint arXiv:2402.11943*, 2024.

Yuzhou Yang, Yangming Zhou, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. Search, examine and early-termination: Fake news detection with annotation-free evidences. *arXiv preprint arXiv:2407.07931*, 2024.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1): 1, 2023.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pp. 2733–2743. ACM, July 2023. doi: 10.1145/3539618.3591879. URL http://dx.doi.org/10.1145/3539618.3591879.

Xin Yuan, Jie Guo, Weidong Qiu, Zheng Huang, and Shujun Li. Support or refute: Analyzing the stance of evidence to detect out-of-context mis-and disinformation. *arXiv preprint arXiv:2311.01766*, 2023.

Fengzhu Zeng, Wenqian Li, Wei Gao, and Yan Pang. Multimodal misinformation detection by learning from synthetic data with multimodal llms. *arXiv preprint arXiv:2409.19656*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.

Yizhou Zhang, Loc Trinh, Defu Cao, Zijun Cui, and Yan Liu. Detecting out-of-context multimodal misinformation with interpretable neural-symbolic model. *arXiv preprint arXiv:2304.07633*, 2023.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. Fact-checking meets fauxtography: Verifying claims about images. *arXiv preprint arXiv:1908.11722*, 2019.

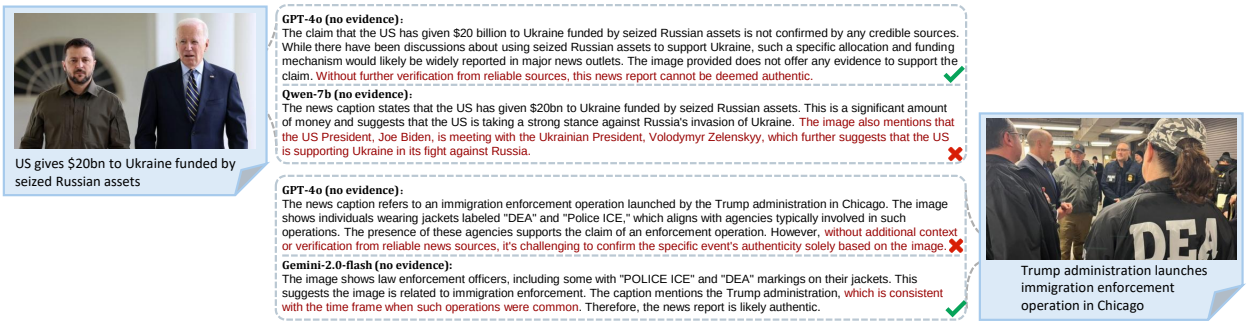

Figure 6: Comparisons of different MLLMs' performance without evidence. Different MLLMs exhibit distinct prediction tendencies in the absence of supporting evidence.

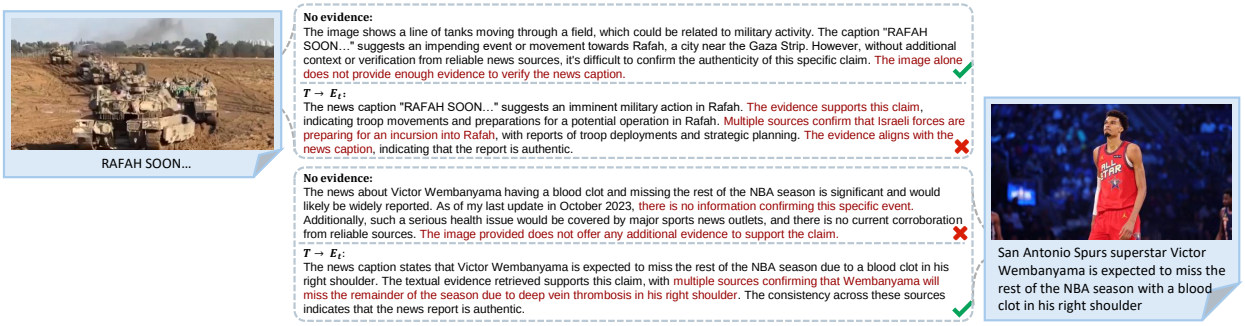

Figure 7: Effectiveness of $T \rightarrow E_t$ strategy on real and fake posts. $T \rightarrow E_t$ is good at finding evidence for real posts. However, it may also leads the model to flip its originally correct prediction, particularly for image OOC misinformations.

A More Analysis of Evidence Retrieval

MLLMs exhibit notable differences in their behavior without evidence. As shown in Fig. 6, GPT shows a conservative tendency in multimodal misinformation detection. Whether the post is fake (as shown in the upper figure) or real (as in the lower figure), GPT tends to classify it as fake when no supporting evidence is available. In contrast, Gemini and Qwen exhibit the opposite behavior: they are more likely to classify the news as real if no clear inconsistency is observed between the image and the caption. This further highlights that relying solely on the model's internal knowledge, without external evidence, is unreliable for misinformation detection.

$T \rightarrow E_t$ (strategy ①) substantially boosts performance, especially for real posts. Two additional examples are shown in the bottom of Fig. 7 and the top of Fig. 8. We then discuss why the accuracy for fake posts does not notably improve, and even declines slightly for GPT-4o. We attribute this to image OOC misinformation, where $T \rightarrow E_t$ provides no information about the image I, and the strong support for T in E_t misleads the model to flip its originally correct prediction of fake. As illustrative examples, two cases are shown in Fig. 9 and at the top of Fig. 7. Without evidence, the model gives a cautious and right answer, while with $T \rightarrow E_t$ supporting the post claim T, it becomes more confident but makes a wrong prediction. Therefore, evidence that directly targets image OOC misinformation serves as an important complement to this evidence such as the example shown on the right side of Fig. 3.

$I \rightarrow E_t$ (strategy ③) is more effective than $T \rightarrow E_i$ (strategy ②) for out-of-context misinformation. Although $T \rightarrow E_i$ shows higher overall accuracy in Table 2, manual inspection reveals that $I \rightarrow E_t$ better detects out-of-context cases. This is because $I \rightarrow E_t$ retrieves webpages directly containing the query image and extracts highly relevant text, while $T \rightarrow E_i$ conducts a fuzzy search based on the caption and often retrieves loosely related images. Textual evidence is also more informative in these cases, since image-based

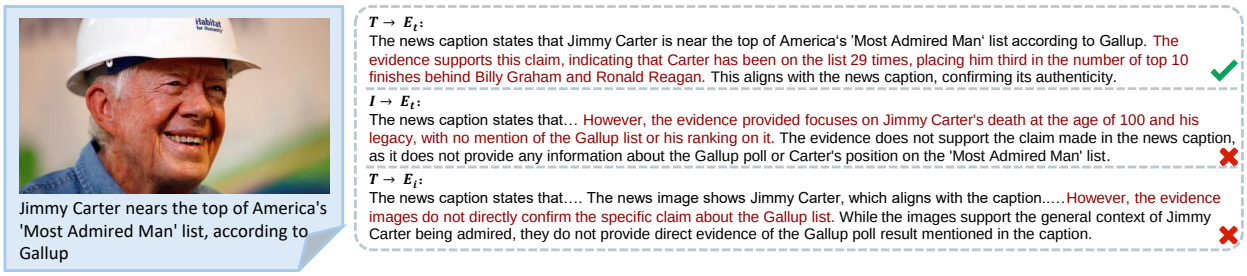

Figure 8: Effectiveness of different evidence types on real posts. $T \rightarrow E_t$ can effectively retrieve relevant evidence for real posts, but cross-modal evidence is less useful in this case.

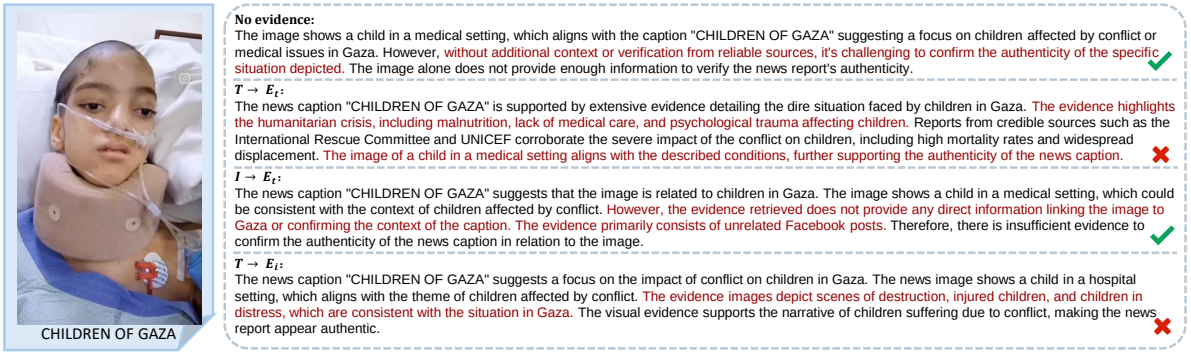

Figure 9: Effectiveness of different evidence types on out-of-context misinformation posts.

comparisons are often limited to coarse features like general scenes or people. These surface-level similarities are usually preserved in out-of-context misinformation, making it hard to detect manipulation through image evidence alone. As shown in Fig. 9, $I \rightarrow E_t$ correctly traces the image to a Facebook user and identifies it as out-of-context misinformation. In contrast, $T \rightarrow E_i$ provides only rough background comparison, failing to precisely detect the misinformation.

DuckDuckGo provides lower-quality evidence than Google. We present one example here in Fig. 11, where DuckDuckGo News Search (strategy ⑧) failed to retrieve any relevant evidence, while Google Search (strategy ①) accurately returned the "direct match" news events that helped the model make the correct inference.

Domain Filter can mitigate evidence noise. Domain filter can improve accuracy by filtering out misleading evidence from low-credibility websites. Here we give an example. As shown in Fig. 10, without the domain filter, there is relevant content in the evidence that supports the claim made in the caption. Therefore, even though the news post caption itself mentions that this might be an "unconfirmed rumor", the model still classifies it as true based on the supporting evidence. However, the domain filter excludes this incorrect evidence, making it easy for the model to make the correct judgment. This shows that incorrect evidence can still have a significant negative impact on detection, even in cases where the correct classification should be straightforward.

LLM-Based Evidence Extraction can mitigate evidence noise. We present an example in Fig. 12. The extracted evidence is more concise, making it easier for the model to perform reasoning.

B More Analysis of Reasoning

Details of reasoning methods:

- **Chain of Thought**: Model outputs an additional rationale in addition to the binary label.

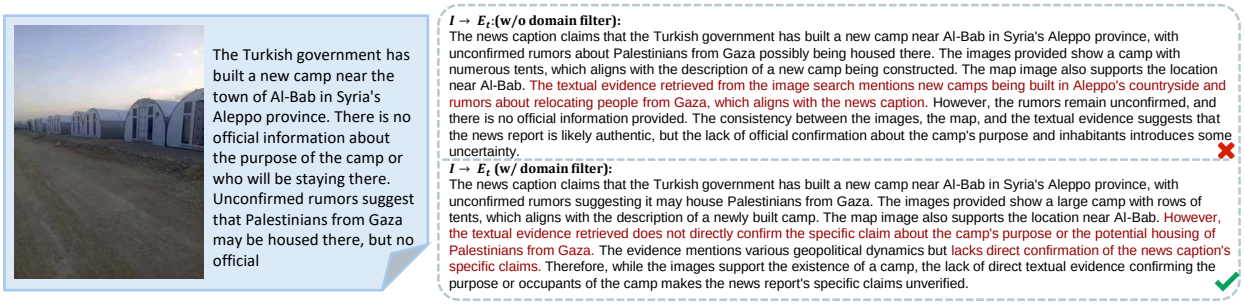

Figure 10: Effectiveness of evidence domain filter.

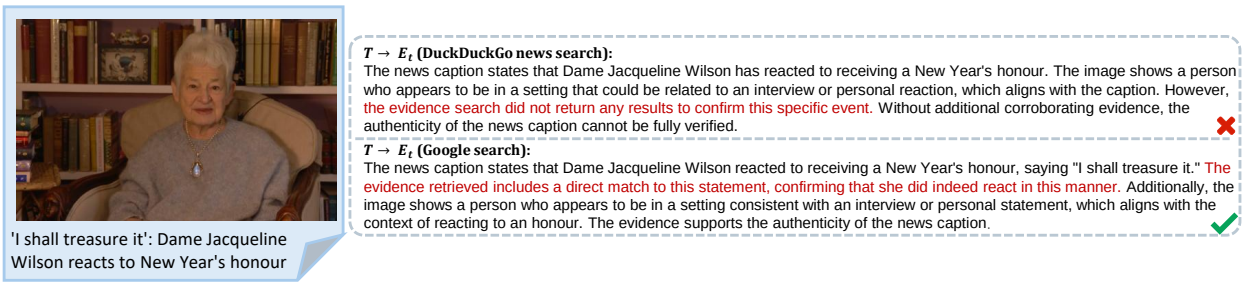

Figure 11: Comparison between Google Search and DuckDuckGo News Search.

- **Prompt Ensembles**: Inspired by Geng et al. (2024), we use a variety of prompts to generate multiple responses, then ask the model to aggregate the responses to get a more robust result.
- **Self Consistency**: Perform multiple rounds of inference and use majority vote to obtain the final result.
- **Multi-Agent Reasoning**: The model may become confused when multiple sources of evidence are provided. Therefore, we invoke the LLM separately for each type of evidence and then summarize all intermediate reasoning processes to produce a final aggregated answer.

Comparison between CoT and multi-agent reasoning. Figure 13 shows the different reasoning paths between CoT reasoning and multi-agent reasoning. In this example, multi-agent reasoning accurately identifies that the image originates from another event by analyzing $I \rightarrow E_t$, leading to the correct classification as image OOC misinformation. However, the CoT reasoning fails to fully utilize each piece of evidence, leading it to overlook the $I \rightarrow E_t$ evidence and resulting in an incorrect inference.

Comparison of reasoning performance across different model sizes. Figure 14 presents the reasoning paths of GPT-4o and GPT-4o-mini. GPT-4o has stronger reasoning capabilities than GPT-4o-mini, which allows it to more precisely recognize the phrase "initially entered a not-guilty plea" in the evidence and therefore make the correct judgement.

C Comparison of Different Misinformation Detection Datasets

Results are shown in Table 10. GPT-4o achieves an accuracy of 0.8 or even 0.9 without using any evidence on other datasets, indicating that it can perform misinformation detection effectively through memorization alone. Moreover, we observe that the improvement brought by evidence is most significant on our dataset. Therefore, our dataset is more suitable for evaluating retrieval-based misinformation detectors and has less evaluation bias compared to real-world misinformation scenarios.

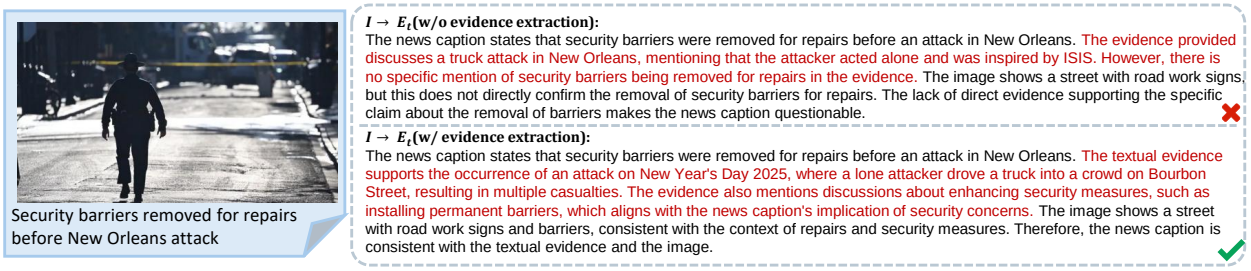

Figure 12: Effectiveness of evidence extraction.

Table 10: Comparison of GPT-4o's performance on different datasets.

Dataset	Evidence	Acc.	R. Acc.	F. Acc.
VERITR	✗	80.1	78	82
	✓	91.9	88.5	95.2
Snopes+Reuters	✗	91.9	93.5	90.3
	✓	96.7	94.6	99
Newsclipping	✗	80.7	88.2	73.2
	✓	89.5	91.2	87.7
Xfacta(Ours)	✗	70.5	51	90
	✓	89.5	98	81

D Example of Detector-Assisted Dataset Expansion

Fig. 15 shows an example of misinformation detection on the newest posts from X using our detector, including the supporting explanations for each prediction to assist human reviewers in verifying the results more efficiently.

E Effect of Other Evidence Aggregation Strategies

In Table 11, All Evidence denotes using all available evidence types jointly, All Evidence + Extraction further applies the proposed evidence extraction module to filter and condense the retrieved evidence, and T $\rightarrow Et$ + T $\rightarrow Ei$ + I $\rightarrow Et$ represents a direct combination of three individual retrieval directions. Using all evidence sources yields strong performance, and evidence extraction further improves accuracy on both real and fake posts. In contrast, directly aggregating multiple retrieval directions results in lower performance, suggesting that naive combination of heterogeneous evidence may introduce noise.

F Effectiveness of OT-based Image Debiasing

We further analyze the effectiveness of the Optimal Transport (OT)–based image debiasing strategy used in our dataset construction. Specifically, we examine whether OT improves (i) feature-level alignment between real and fake images, and (ii) reduces image-only discriminability. Table 12 shows that OT-based selection substantially increases the cosine similarity between real and fake image features (from 0.19 to 0.38), indicating improved alignment in the visual feature space. This suggests that OT effectively reduces distributional discrepancies that may otherwise introduce spurious visual cues. Consistent with this observation, Table 13 shows that after OT selection, GPT-4o becomes less biased in image-only real vs. fake classification. In particular, the prediction distribution for images from real posts moves closer to random guessing, indicating that visually exploitable shortcuts are reduced. Together, these results demonstrate that OT-based image debiasing not only aligns feature distributions but also mitigates image-induced shortcut learning, leading to a more reliable evaluation of multimodal misinformation detection models.

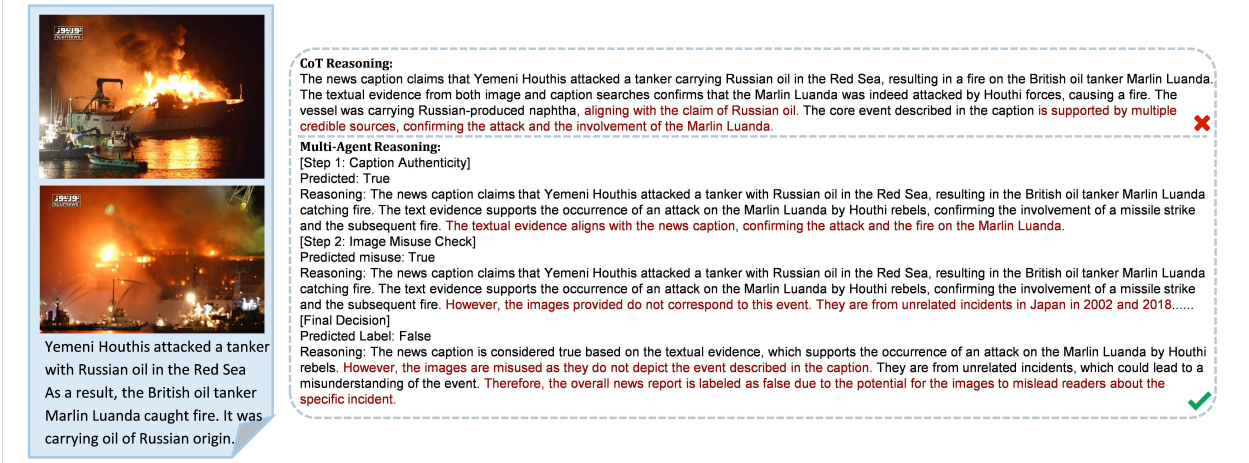

Figure 13: Comparison between CoT reasoning and Multi-Agent reasoning.

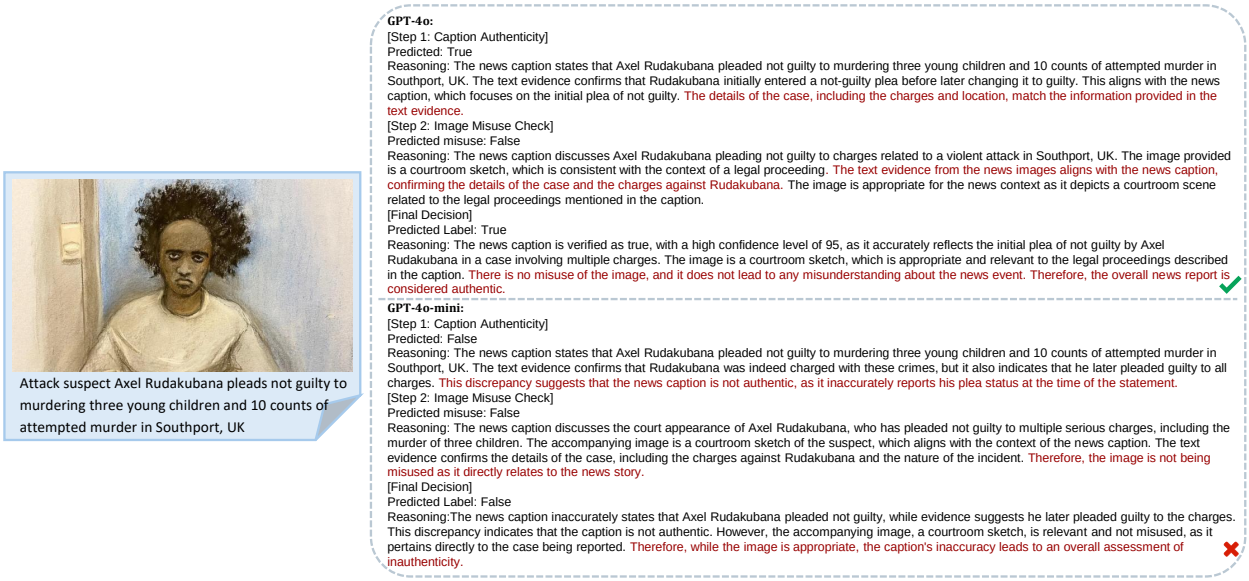

Figure 14: Comparison between GPT-4o and GPT-4o-mini.

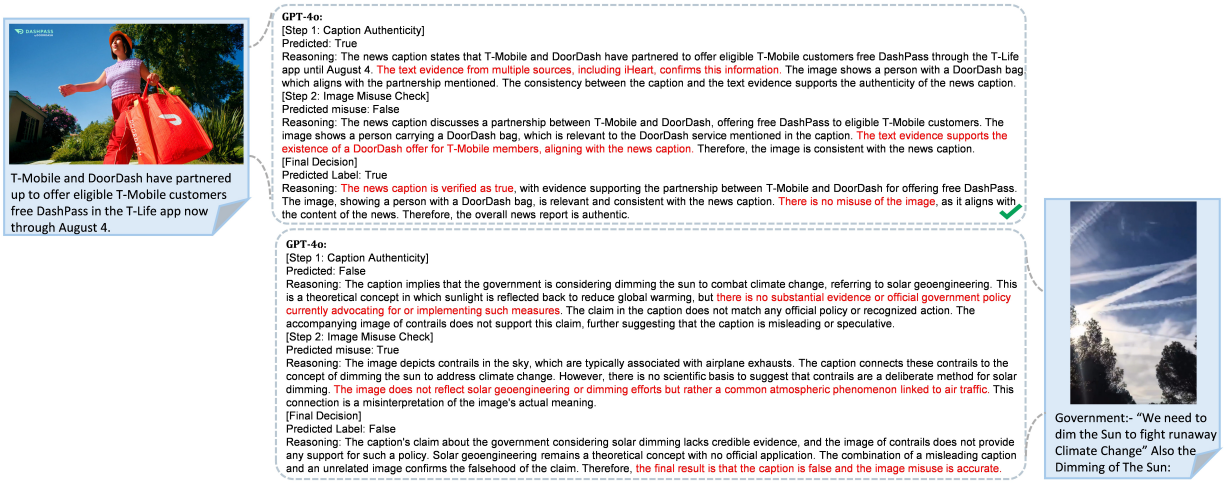

Figure 15: Example of misinformation detection on newest posts from X using the detector's inference

Table 11: Performance of combined evidence strategies.

Experiment Setup	**Acc (Real)**	**Acc (Fake)**	**Acc (Overall)**
All Evidence	96.7	80.0	88.3
All Evidence + Extraction	98.3	84.2	91.3
$\mathrm{T} \rightarrow E_t + \mathrm{T} \rightarrow E_i + \mathrm{I} \rightarrow E_t$	81.7	69.2	75.4

Table 12: Feature-level alignment between real and fake images measured by cosine similarity using SigLip features.

Setting	**Cosine Similarity**
Random Selection	0.19
OT Selection	0.38

Table 13: Image-only classification results by GPT-4o under different image selection strategies.

Setting	**Images from real posts**	**Images from fake posts**
Random Selection	47% real / 53% fake	37% real / 63% fake
OT Selection	41% real / 59% fake	37% real / 63% fake