

000 001 XFACTA: CONTEMPORARY, REAL-WORLD DATASET 002 AND EVALUATION FOR MULTIMODAL MISINFORMA- 003 TION DETECTION WITH MULTIMODAL LLMs 004 005

006 **Anonymous authors**
007 Paper under double-blind review
008
009
010

011 ABSTRACT 012

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018 On the dataset side, existing benchmarks either contain *outdated* events, leading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

031 1 INTRODUCTION 032

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

040 The emergence of multimodal large language models (MLLMs), with strong reasoning capabilities
041 across both text and images, offers a promising direction for detecting multimodal misinformation.
042 Recent studies have begun to explore this potential. Some methods (Qi et al., 2024; Liu et al.,
043 2024a; Zeng et al., 2024; Shalabi et al., 2024) fine-tune a general-purpose MLLM on specific mis-
044 information datasets to create task-specific models. Other approaches (Khaliq et al., 2024; Xuan
045 et al., 2024; Liu et al., 2024b; Geng et al., 2024) adopt a zero-shot setting and rely on more power-
046 ful models such as GPT-4 or Gemini, which achieve better performance on existing misinformation
047 datasets. In general, the existing MLLM-based misinformation detectors mimic human verification
048 processes, which involves two main steps: *evidence retrieval*, where external information is retrieved
049 from Internet to serve as evidence, then *reasoning*, where the news post and the retrieved evidence
050 are systematically analyzed and combined to make final judgment.

051 Despite the promising results reported in these studies, it remains unclear exactly where the bottle-
052 neck of existing MLLM-based misinformation detection methods lies (evidence retrieval *v.s.* rea-
053 soning), hindering 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’ train-

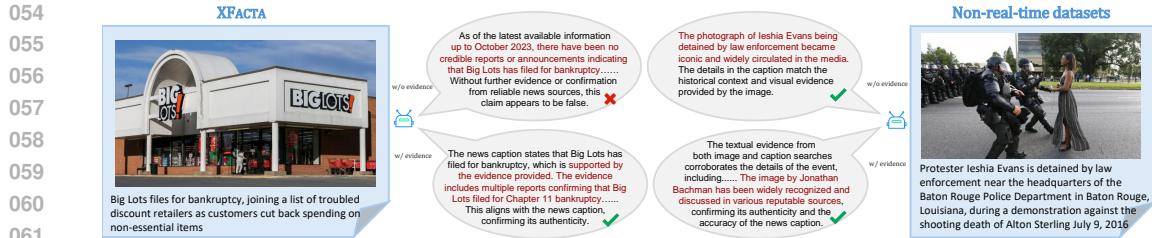


Figure 1: **Left:** An example from our dataset, where the MLLM (GPT-4o) must rely on evidence to judge real or fake. **Right:** An example from non-real-time datasets, where evidence matters less. Evaluating MLLM-based misinformation detectors on XFACTA introduces less evaluation bias.

ing 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) 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.

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;

108
109
110
111
112

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.

113
114
115
116
117
118
119
120
121

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
NewsCLIPpings (Luo et al., 2021)	✓	✗	✗	✗	816,922	816,922
Fakedit (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
MocheG (Yao et al., 2023)	✓	✗	✓	✓	5,144	5,855
MediaEval (Boididou et al., 2016)	✓	✗	✓	✓	292	410
VERITE (Papadopoulos et al., 2023)	✓	✗	✓	✓	338	662
Post-4V (Geng et al., 2024)	✓	✓	✓	✓	81	105
XFACTA (Ours)	✓	✓	✓	✓	1,200	1,200

122
123124
125 Khanam et al., 2021). The rise of social media has led to increasing attention on multimodal mis-
126 information detection, along with the release of various datasets (Nakamura et al., 2019; Zlatkova
127 et al., 2019; Yao et al., 2023; Boididou et al., 2016; Papadopoulos et al., 2023). However, real-world
128 misinformation datasets are typically either small in size or suffer from noisy annotations. There-
129 fore, some other works (Luo et al., 2021; Chakraborty et al., 2023; Liu et al., 2024b; Shao et al.,
130 2023; Aneja et al., 2021) leverage heuristic rules or AI models to synthesize datasets for mis-
131 information detection. As a consequence of such synthesis, these datasets fail to capture real-world
132 misinformation creators’ complex patterns and strategies. In addition, all of the above datasets are
133 not contemporary and often overlap with the training data of MLLMs, which prevents a fair and
134 robust evaluation of MLLM-based misinformation detectors. Post-4V (Geng et al., 2024) addresses
135 this by collecting more recent examples, but its size is very limited, and data collection and process-
136 ing details are underdocumented, making it less suitable as a widely accepted baseline. In contrast,
137 our XFACTA dataset ensures both contemporary and real-world characteristics, while maintaining a
138 moderate scale that is sufficient to evaluate MLLMs in a zero-shot setting. In addition, our dataset
139 provides detailed journalist evidence for fake news, which can help validate the reasoning paths of
140 detection models. A multi-dimensional comparison across different datasets can be found in Table 1.
141142 **Models:** Some traditional multimodal misinformation detectors Abdelnabi et al. (2022); Yuan et al.
143 (2023); Brahma et al. (2023a); Aneja et al. (2023); Mu et al. (2023); Zhang et al. (2023); Brahma
144 et al. (2023b); Yang et al. (2024) are trained and evaluated on specific datasets, such as the com-
145 monly used NewsCLIPpings dataset (Luo et al., 2021). With the emergence of open-source MLLMs,
146 several recent works (Qi et al., 2024; Liu et al., 2024a; Zeng et al., 2024; Shalabi et al., 2024) have
147 adopted a different approach by fine-tuning a pretrained MLLM on misinformation datasets, which
148 achieves better performance. However, these methods often carry biases specific to their training
149 data, which are not robust to new, more sophisticated misinformation emerging on social platforms.
150 Therefore, several studies have explored more powerful closed-source MLLMs and have achieved
151 better results. However, these models are either claimed evidence-free (Geng et al., 2024), or eval-
152 uated on less updated or real-world datasets (Khalil et al., 2024; Xuan et al., 2024; Liu et al., 2024b;
153 Jin et al., 2024), raising concerns about their effectiveness when deployed on evolving social media.154
155

3 OUR XFACTA DATASET

156 **Multimodal Misinformation Detection** refers to assessing the authenticity of a news post that in-
157 cludes both supporting images and text. Formally, given a set of supporting images $I = \{I_1, \dots, I_n\}$
158 and a text claim T , this task is to determine whether the post $\mathcal{P} = (I, T)$ is real or fake.159 The supporting images I can make the text claim T seem more believable, even if they are unrelated
160 or misleading, which makes detection much harder than in the unimodal setting. Therefore, most
161 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.

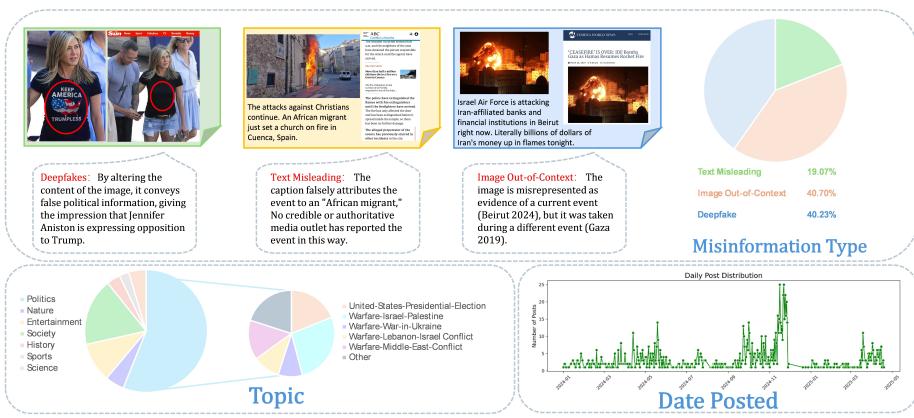


Figure 2: Examples and distribution of misinformation types, topics, and posting dates in XFACTA.

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: the previous step focuses more on aligning the text claims T , here we address the alignment of images I . We use SigLip (Zhai et al., 2023) to extract image features and apply the Optimal Transport algorithm (Genevay et al., 2016) to select real posts whose image feature distribution most closely matches that of the fake posts. This alignment helps minimize the bias caused by visual differences, ensuring the evaluation accurately reflects the model’s true capability in detecting misinformation from both textual and visual perspectives.

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 2400 data points, including 1200 real posts and 1200 fake posts. For the convenience of model development, we randomly selected 120 real and 120 fake posts as the Dev set, while the remaining 2160 posts was used as the Test set. As shown in the bottom right corner of Fig. 2, all data are collected from January 2024 to April 2025, with a majority of data collected after September 2024. This contemporary nature ensures the dataset reflects emerging patterns and evolving characteristics of both real and fake news.

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.

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

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

227 By annotating each fake post with these finer-grained misinformation labels, we achieve a more nu-
 228anced understanding of the characteristics of multimodal misinformation, and enable a more detailed
 229 analysis of a misinformation detector’s performance across different misinformation types.

231 4 HOW TO BUILD A GOOD MLLM-BASED MISINFORMATION DETECTOR?

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

238 4.1 ANALYSIS OF EVIDENCE RETRIEVAL

240 4.1.1 EXPERIMENT SETUP

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

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

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

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

266 To evaluate the impact of each evidence type, we first run the model without evidence, relying only
 267 on an MLLM’s internal knowledge. Then, we add each of the eight evidence types separately and
 268 compare the results against the no-evidence baseline and with each other. We use Chain-of-Thought

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

270

271
272
Table 2: Comparison of MLLM Performance with varying ev-
idence 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

279

280

281

282

283

284

285

286

287

288

289

290

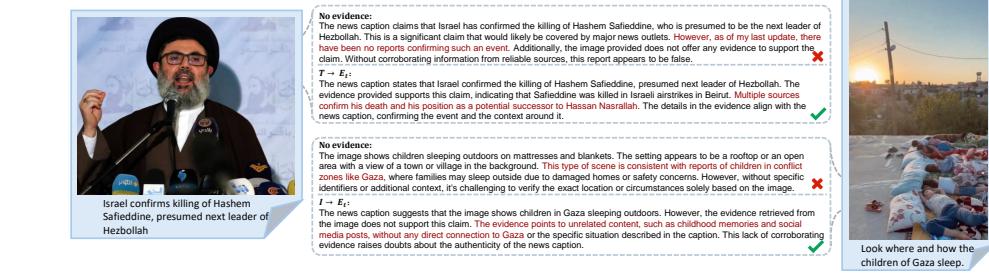
291

292

293

294

295

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

(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. 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

324
325 Table 4: Comparison of each evidence retrieval strategies across misinformation types with GPT-4o.
326

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

333
334 Table 5: Comparison of MLLM Performance with various reasoning methods on the Dev set.
335

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-step Reasoning	91.3	91.7	90.8	81.3	90	72.5	62.1	78.4	45.9

344
345 like general scenes or people. These superficial similarities are usually preserved in out-of-context
346 misinformation, making it hard to detect manipulation through image-type evidence. Point 5 further
347 analyzes strategy ③ on fake posts.

348
349 **4. LLM-generated queries (strategies ④ and ⑤) are less effective than direct caption searches.**
350 In most cases, an LLM is not able to generate highly targeted queries; most of them are simply
351 paraphrases of the original caption. Searching with such paraphrased versions is thus less accurate
352 than directly using the caption T itself to retrieve evidence. In certain cases of fake posts, if the
353 questions or doubts raised by the LLM fail to target the actual reason why the post is fake, the
354 retrieved evidence can even lead the model to confidently make an incorrect judgment, as further
355 analyzed in the next point.

356
357 **5. Across different misinformation types, $I \rightarrow E_t$ (strategy ③) provides consistently informative**
358 **evidence especially for identifying fake posts.** Unlike earlier analyses based on the overall
359 accuracy, analyzing fine-grained misinformation types for fake posts demands careful consideration
360 beyond accuracy alone. GPT-4o tends to conservatively classify ambiguous posts as fake even with-
361 out additional evidence, and this can inflate the accuracy of fake posts. Hence, average confidence
362 scores become essential because they indicate whether the retrieved evidence provides clear and
363 informative knowledge that truly helps the model’s judgment. As shown in Table 4, strategy ③
364 not only achieves high accuracy but also consistently maintains high confidence across Deepfakes
365 and Image OOC categories. Although Query $\rightarrow E_t$ (strategy ⑤) shows slightly superior combined
366 performance in the Text Misleading category, it causes the model to make highly confident but in-
367 correct predictions in Image OOC cases, significantly reducing its overall utility. Therefore, $I \rightarrow E_t$
368 remains the optimal evidence retrieval strategy across various misinformation types.

369
370 **6. DuckDuckGo provides lower-quality evidence than Google (strategies ⑥-⑧).** Experiments
371 indicate that evidence retrieved using DuckDuckGo consistently yields lower performance compared
372 to Google Search. Additionally, $T \rightarrow E_{news}$ did not produce the expected improvements.

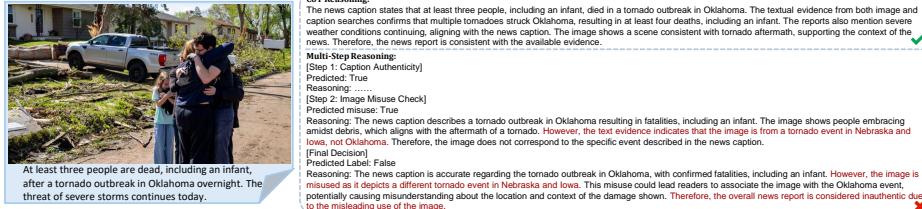
373
374 **7. Domain Filter can mitigate evidence noise.** As shown in Table 3, domain filter can improve
375 accuracy in general, suggesting that evidence from low-credibility websites is indeed noisy and
376 potentially misleading.

377
378 **8. LLM-Based Evidence Extraction can mitigate evidence noise.** We inspected the extraction
379 results and found that the LLM can successfully retain the key information needed to detect mis-
380 information and filter out some irrelevant evidence, which leads to improved detection accuracy as
381 shown in Table 3, especially for $I \rightarrow E_t$. However, it is important to note that evidence extraction
382 introduces a huge token overhead.

378
379
380
381
382
383
384
385
386
387

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

388
389
390
391
392
393
394
395396
397
398
Figure 4: Multi-step reasoning can detect image “misuse” in the news post from CNN. We believe
399
400 this “overly strict” behavior is actually beneficial for reliable misinformation detection.401
402

4.2 ANALYSIS OF REASONING

403
404
405
406
407

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-step Reasoning**. See the appendix for additional details.

408
409

4.2.2 RESULTS AND ANALYSIS

410
411
412
413
414

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

415
416
417

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.

418

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.

419
420
421
422
423
424
425
426

3. For GPT-4o, multi-step 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-step 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-step 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.

427
428
429

5 FURTHER EVALUATIONS ON XFACTA

430
431

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

432 types following Section 4.2 and use the multi-step reasoning strategy. Results are shown in Table 6.
 433 For the same model architecture, larger models always achieve higher accuracy.
 434

435 **Comparison of Existing Multimodal Misinformation Detection Methods.** We perform a horizontal
 436 comparison with existing multimodal misinformation approaches using our Test set, in-
 437 cluding models trained from scratch: SENs (Yuan et al., 2023), Mocheg (Yao et al., 2023), and
 438 HAMMER (Shao et al., 2023); methods fine-tuning MLLMs: Sniffer (Qi et al., 2024), and zero-
 439 shot methods using closed-source MLLMs: MMFakeBench (Liu et al., 2024b), LEMMA (Xuan
 440 et al., 2024). Results in Table 7 show that specialist methods that trained on a specific dataset suf-
 441 fer from severe generalization issues. In contrast, models that use GPT-4o demonstrate relatively
 442 good performance. Based the systematic analysis of evidence retrieval and reasoning strategies, our
 443 method outperforms these models, establishing SOTA accuracy on XFACTA.

444 **Comparison of Similar Datasets.** We examine the dependency on evidence across similar datasets.
 445 To achieve this, we select a subset from (Papadopoulos et al., 2023), Snopes+Reuters Zlatkova et al.
 446 (2019), and NewsCLIPpings (Luo et al., 2021) dataset, respectively. We use the evidence types
 447 following Section 4.2 and apply CoT reasoning with GPT-4o. We report the results in Table 8 in
 448 the Appendix. Notably, for our XFACTA dataset, GPT-4o does not work well without any evidence,
 449 confirming the need for a contemporary, real-world benchmark.

450 **Detector Effectiveness on More Recent and Out-of-Distribution Data.** We evaluate whether the
 451 misinformation detector trained on the original XFACTA dataset remains effective when applied to
 452 more recent and out-of-distribution social media content. First, we choose Snopes as the testbed
 453 since, compared to X, it is out-of-distribution. Moreover, the website provides real/fake annotations
 454 which are provided by professional journalists, thus can serve as a reference for evaluating the
 455 performance of the detector. We collect 1,200 fact-checked news items (600 real and 600 fake) from
 456 Snopes between July 2024 and July 2025, which are more recent than original XFACT dataset. Our
 457 resulting detector achieves an overall accuracy of 89.2% on this dataset (85.5% on true and 93.0%
 458 on false), showing that the model works well on both newest and out-of-distribution data.

459 6 CLOSING THE LOOP: DETECTOR-ASSISTED DATASET EXPANSION

460 In this section, we demonstrate how our detector, which has been validated on the original XFACTA
 461 dataset, can be effectively used to support dataset expansion. Previous experiments show that the
 462 detector maintains stable performance on both more recent and out-of-distribution data, suggesting
 463 that it generalizes well to continuously emerging, previously unseen content. Therefore, it can be
 464 integrated into the dataset collection pipeline to assist human reviewers in verifying misinformation,
 465 enabling a detection-in-the-loop framework that accelerates and scales up the data curation process.
 466

467 To verify this idea, we conduct a case study as a proof of concept. This time, we do not rely on
 468 journalist-flagged posts or posts from official news accounts for real/fake references. Instead, we
 469 select several accounts that regularly post about trending or controversial topics and that have a
 470 sizable follower base. We crawl and identify 500 posts between June 2025 and July 2025 from
 471 these accounts. Among them, 265 posts are identified by our detector as fake and 235 as real. For
 472 each prediction, the detector also generates an explanation to support its decision. The explanations
 473 can assist human reviewers in verifying the predictions more efficiently and deciding whether to
 474 incorporate the posts into XFACTA. An example can be found in Appendix. The additional dataset
 475 has been released alongside the main XFACTA dataset as part of our public release.
 476

477 7 CONCLUSION

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

486 ETHICS STATEMENT
487488 Our research adheres to the guidelines set forth by the Twitter Developer Terms². We ensure that our
489 data collection and use comply with these terms, including the appropriate use of the Twitter API.
490 While we plan to release our dataset for research purposes, we will do so in a manner that adheres
491 to all applicable rules and guidelines.492 Our study focuses on detecting multimodal misinformation, a significant issue in the digital age.
493 By identifying and mitigating the spread of misinformation, our work contributes positively to the
494 integrity of information on the web. Since our dataset consists of internet fake news posts, some
495 posts may contain offensive content. However, the positive contributions of our research in reducing
496 misinformation far outweigh the potential negatives.
497498 REFERENCES
499500 Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. Open-domain, content-based, multi-modal fact-
501 checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF confer-
502 ence on computer vision and pattern recognition*, pp. 14940–14949, 2022.503 Shivangi Aneja, Chris Bregler, and Matthias Nießner. Cosmos: Catching out-of-context misinfor-
504 mation with self-supervised learning, 2021. URL <https://arxiv.org/abs/2101.06278>.
505506 Shivangi Aneja, Chris Bregler, and Matthias Nießner. Cosmos: catching out-of-context image mis-
507 use using self-supervised learning. In *Proceedings of the AAAI conference on artificial intelli-
508 gence*, volume 37, pp. 14084–14092, 2023.509 Christina Boididou, Symeon Papadopoulos, Duc Tien Dang Nguyen, G. Boato, Michael Riegler,
510 Andreas Petlund, and Ioannis Kompatsiaris. Verifying multimedia use at mediaeval 2016. 10
511 2016.
512513 Debarshi Brahma, Amartya Bhattacharya, Suraj Nagaje Mahadev, Anmol Asati, Vikas Verma, and
514 Soma Biswas. Dpod: Domain-specific prompt tuning for multimodal fake news detection. *arXiv
515 preprint arXiv:2311.16496*, 2023a.516 Debarshi Brahma, Amartya Bhattacharya, Suraj Nagaje Mahadev, Anmol Asati, Vikas Verma, and
517 Soma Biswas. Leveraging out-of-domain data for domain-specific prompt tuning in multi-modal
518 fake news detection. *arXiv preprint arXiv:2311.16496*, 2023b.
519520 Megha Chakraborty, Khusbu Pahwa, Anku Rani, Adarsh Mahor, Aditya Pakala, Arghya Sarkar,
521 Harshit Dave, Ishan Paul, Janvita Reddy, Preethi Gurumurthy, et al. Factify3m: A benchmark for
522 multimodal fact verification with explainability through 5w question-answering. *arXiv preprint
523 arXiv:2306.05523*, 2023.524 Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for
525 large-scale optimal transport. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Gar-
526 nett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Asso-
527 ciates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/2a27b8144ac02f67687f76782a3b5d8f-Paper.pdf.
528529 Jiahui Geng, Yova Kementchedjhieva, Preslav Nakov, and Iryna Gurevych. Multimodal large lan-
530 guage models to support real-world fact-checking, 2024. URL <https://arxiv.org/abs/2403.03627>.
531533 Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. A richly an-
534 notated corpus for different tasks in automated fact-checking. *arXiv preprint arXiv:1911.01214*,
535 2019.536 Yiqiao Jin, Minje Choi, Gaurav Verma, Jindong Wang, and Srijan Kumar. Mm-soc: Benchmarking
537 multimodal large language models in social media platforms. *arXiv preprint arXiv:2402.14154*,
538 2024.
539²<https://developer.x.com/en/developer-terms/more-on-restricted-use-cases>

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

543

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

547

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

552

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

556

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

559

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

563

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

566

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

570

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

574

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

578

579 Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and grounding multi-modal media manipulation,
 580 2023. URL <https://arxiv.org/abs/2304.02556>.

581

582 Snopes. URL <https://www.snopes.com/>.

583

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

587

588 James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-
 589 scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent
 590 (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association
 591 for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp.
 592 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi:
 593 10.18653/v1/N18-1074. URL <https://aclanthology.org/N18-1074/>.

594

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

598

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

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

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

602

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

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

607

608 X/Twitter. URL <https://x.com/>.

609

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

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

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

619

620 Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. End-to-end multi-
 621 modal fact-checking and explanation generation: A challenging dataset and models. In *Proce-
 622 dings of the 46th International ACM SIGIR Conference on Research and Development in Infor-
 623 mation Retrieval*, SIGIR ’23, pp. 2733–2743. ACM, July 2023. doi: 10.1145/3539618.3591879.
 624 URL <http://dx.doi.org/10.1145/3539618.3591879>.

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

628

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

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

633

634 Yizhou Zhang, Loc Trinh, Defu Cao, Zijun Cui, and Yan Liu. Detecting out-of-context multimodal
 635 misinformation with interpretable neural-symbolic model. *arXiv preprint arXiv:2304.07633*,
 636 2023.

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

639

640

641

642

643

644

645

646

647

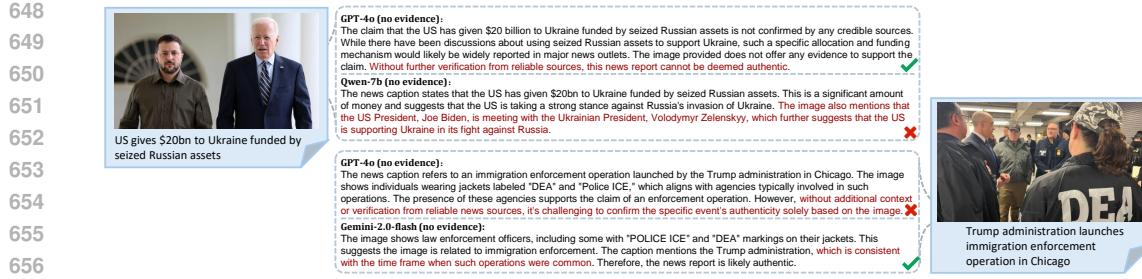


Figure 5: Comparisons of different MLLMs’ performance without evidence. Different MLLMs exhibit distinct prediction tendencies in the absence of supporting evidence.

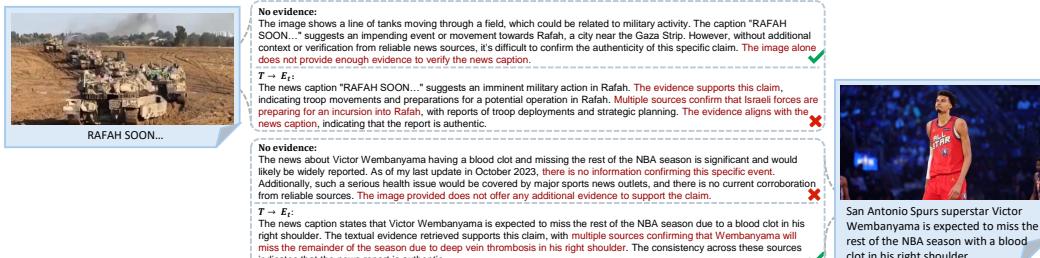


Figure 6: Effectiveness of $T \rightarrow E_t$ strategy on real and fake posts. $T \rightarrow E_t$ is good at finding evidence for real posts. However, it may also lead the model to flip its originally correct prediction, particularly for image OOC misinformations.

A MORE ANALYSIS OF EVIDENCE RETRIEVAL

MLLMs exhibit notable differences in their behavior without evidence. As shown in Fig. 5, 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. 6 and the top of Fig. 7. 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. 8 and at the top of Fig. 6. 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 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. 8, $I \rightarrow E_t$ correctly traces the image to a Facebook user and identifies it as out-of-context mis-

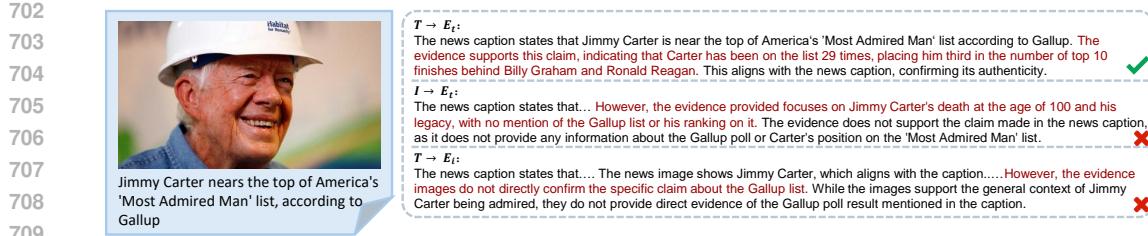


Figure 7: Effectiveness of different evidence types on real posts. $T \rightarrow E_t$ can effectively retrieve relevant evidence for real posts, but cross-modal evidence is less useful in this case.

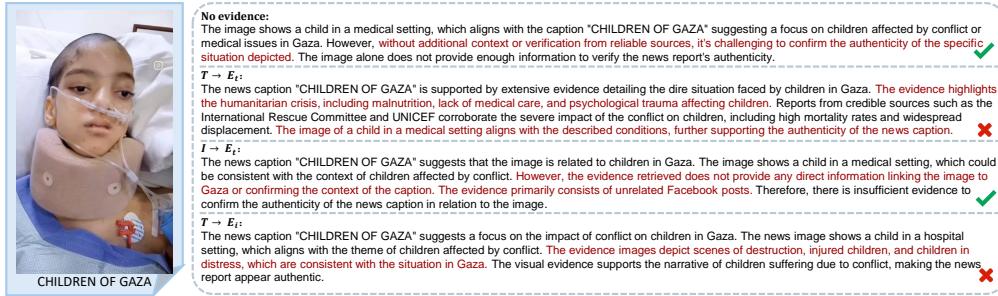


Figure 8: Effectiveness of different evidence types on out-of-context misinformation posts.

information. 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. 10, 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. 9, 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. 11. 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.
- **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-step 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.

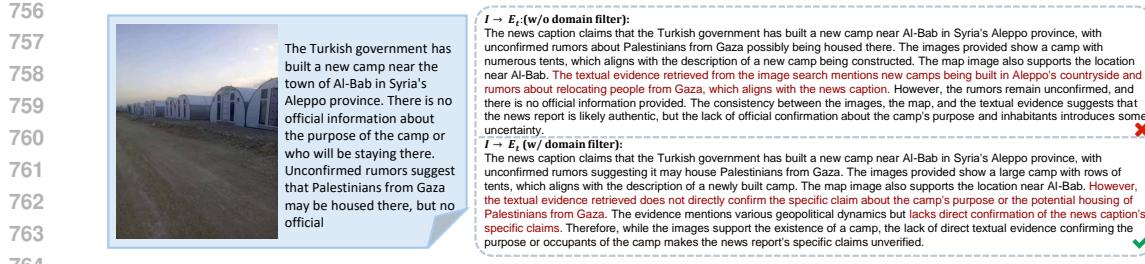


Figure 9: Effectiveness of evidence domain filter.

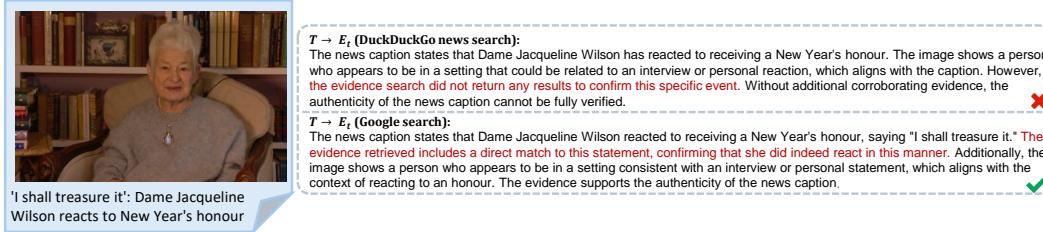


Figure 10: Comparison between Google Search and DuckDuckGo News Search.

Comparison between CoT and multi-step reasoning. Figure 12 shows the different reasoning paths between CoT reasoning and multi-step reasoning. In this example, multi-step 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 13 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 8. 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.

D EXAMPLE OF DETECTOR-ASSISTED DATASET EXPANSION

Fig. 14 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.

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889



T-Mobile and DoorDash have partnered up to offer eligible T-Mobile customers free DashPass in the T-Life app now through August 4.

GPT-4b:

[Step 1: Caption Authenticity]

Predicted: True

Reasoning: The news caption states that T-Mobile and DoorDash have partnered to offer eligible T-Mobile customers free DashPass through the T-Life app until August 4. The text evidence from multiple sources, including iHeart, supports this information. The image shows a person with a DoorDash bag, which is relevant to the DoorDash service mentioned in the caption. The text evidence supports the authenticity of the news caption.

[Step 2: Image Misuse Check]

Predicted misuse: False

Reasoning: The news caption discusses a partnership between T-Mobile and DoorDash, offering free DashPass to eligible T-Mobile customers. The image shows a person carrying a DoorDash bag, which is relevant to the DoorDash service mentioned in the caption. The text evidence supports the authenticity of the news caption.

[Final Decision]

Predicted Label: True

Reasoning: The news caption is verified as true, with evidence supporting the partnership between T-Mobile and DoorDash for offering free DashPass.

The image shows a person with a DoorDash bag, is relevant and consistent with the news caption. There is no misuse of the image, as it aligns with the content of the news. Therefore, the overall news report is authentic.

GPT-4e:

[Step 1: Caption Authenticity]

Predicted: False

Reasoning: The caption implies that the government is considering dimming the sun to combat climate change, referring to solar geoengineering. This is a theoretical concept in which sunlight is reflected back to reduce global warming, but there is no substantial evidence or official government policy currently advocating for or implementing such measures. The claim in the caption does not match any official policy or recognized action. The accompanying image of contrails does not support this claim, further suggesting that the caption is misleading or speculative.

[Step 2: Image Misuse Check]

Predicted: True

Reasoning: The image depicts contrails in the sky, which are typically associated with airplane exhausts. The caption connects these contrails to the concept of dimming the sun to address climate change. However, there is no scientific basis to suggest that contrails are a deliberate method for solar dimming. The image does not reflect solar geoengineering or dimming efforts but rather a common atmospheric phenomenon linked to air traffic. This connection is a misinterpretation of the image's actual meaning.

[Final Decision]

Predicted Label: False

Reasoning: The caption's claim about the government considering solar dimming lacks credible evidence, and the image of contrails does not provide any support for such a policy. Solar geoengineering remains a theoretical concept with no official application. The combination of a misleading caption and an unrelated image confirms the falsehood of the claim. Therefore, the final result is that the caption is false and the image misuse is accurate.



Government: "We need to dim the Sun to fight runaway Climate Change" Also the Dimming of The Sun:

Figure 14: Example of misinformation detection on newest posts from X using the detector's inference

890

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917