Fact-Checking Multimodal Misinformation Using Multimodal Large Language Models

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) have emerged as important tools for analyzing and understanding images as well as claims about them; yet, their effectiveness and limitations for multimodal fact-checking are not well understood. Here we aim to bridge this gap. We conduct extensive experiments to benchmark the fact-checking capabilities of various open-source models of different sizes and types, as well as the leading proprietary model GPT-4o. Our findings reveal that GPT-4o consistently outperforms other models with and without evidence. With access to retrieved evidence, open-source models significantly narrow the performance gap but still lag behind GPT-4o due to their inability to effectively leverage relevant evidence. Given this observation, we implement a knowledge transfer framework that utilizes rationales from GPT-4o to instill better reasoning capabilities in open-source models, particularly in the effective use of evidence. Experimental results demonstrate the efficacy of our approach, showing significant improvements in fact-checking tasks. Our data and code will be released upon publication.

1 Introduction

Online misinformation, and in particularly multimodal misinformation that combines text, images, videos, and other types of media, poses a major challenge (Akhtar et al., 2023). Textual claims are often accompanied by manipulated or out-of-context (OOC) visual components as a way to bolster false narratives (Huh et al., 2018; Luo et al., 2021; Yao et al., 2023). Effective fact-checking, therefore, requires tools that can adeptly handle multiple modalities.

Large Language Models (LLMs) store vast amounts of information beyond human knowledge and provide more user-friendly access than traditional search engines (Zhao et al., 2023; Hu et al., 2024b). As such, LLMs can serve as a powerful tool for human fact-checkers, who frequently seek additional factual knowledge to verify claims (Nakov et al., 2021). Indeed, there is growing interest in leveraging LLMs for fact-checking (Chen et al., 2022a; Cheung and Lam, 2023; Li et al., 2023b). However, existing methods often face limitations when dealing with multimodal information, as they require OCR or image captioning models to extract relevant details from images (Wang et al., 2024). Recent advancements in multimodal LLMs (MLLMs) offer promising opportunities for progress in the field. Pretraining and instruction tuning enable them to analyze images across tasks, providing valuable insights like detecting fake images and delivering image-related knowledge akin to knowledge bases (Hu et al., 2024a; Romero et al., 2024; Keita et al., 2024). Furthermore, these models can flexibly incorporate external retrieved knowledge, leveraging document retrieval based on both visual and textual claims (Chen et al., 2022b).

However, there is a lack of empirical work assessing the performance of current models on fact-checking tasks. Given the limited accessibility and the high cost of proprietary models like GPT-4o, the potential of inexpensive open-source models as viable alternatives remains an open question. Motivated by this, we conduct an extensive empirical study to measure the performance of various models, both relying solely on parametric knowledge and with the help of external knowledge. Our experiments are based on publicly available datasets: *Fauxtography* (Zlatkova et al., 2019), sourced from debunking articles, and MR^2 (Hu et al., 2023), sourced from Twitter posts. Additionally, we introduce a new dataset, *LatestSnopes*, to mitigate potential data leakage and provide a more robust evaluation. The results show that GPT-4o substantially outperforms open-source models in evidence-free scenarios, demonstrating better world knowledge and superior reasoning abilities. Meanwhile, open-

source models from the InternVL2 (Chen et al., 2023) and LLaVA-1.6 (Liu et al., 2024) series perform only marginally better than random guessing without evidence. When provided with evidence, open-source models narrow the performance gap. Yet, upon inspecting the models' rationales, we find that they sometimes struggle to effectively utilize the evidence, ignoring clear textual clues, relying solely on image analysis, or focusing disproportionately on irrelevant details in the claims—issues that are rarely observed in GPT-4o.

Building on these observations, we propose a knowledge transfer framework to enhance the performance of open-source models in fact-checking. In particular, we employ GPT-4o as a teacher, collecting its rationales for various claims and evidence to create an instruction-tuning dataset to improve the open-source models' ability to leverage evidence. Unlike previous methods (Qi et al., 2024; Wang et al., 2024), which rely on text-only models to generate rationales from retrieved textual evidence and ground truth for instruction-tuning datasets and fine-tune multimodal models with a different instruction template, our approach eliminates inconsistencies in instructions. By using a consistent prompt for data annotation, we filter out incorrect predictions while retaining examples where GPT-4o identifies insufficient evidence. This approach enables the model to more effectively detect and articulate gaps in the evidence.

Our contributions can be summarized as follows:

- We conduct a comprehensive empirical evaluation of open-source and proprietary MLLMs across multimodal fact-checking datasets, both with and without retrieved evidence, addressing a gap in the current research.

- Our analysis reveals that external evidence can help address knowledge limitations and ambiguous information. However, open-source models still struggle to effectively leverage this evidence, even when provided with accurate and relevant information.

- We propose a knowledge transfer framework that harnesses GPT-4o's superior evidence utilization and reasoning capabilities to enhance open-source models, with extensive experiments demonstrating that open-source models can approach proprietary model performance, surpassing previous methods.

2 Related Work

Multimodal Fact-Checking Multimodal misinformation detection has attracted significant attention due to the growing threat it poses to the integrity of information dissemination. Content-based approaches primarily assess the consistency between modalities, such as image and text. In this context, various methods have been explored, including multimodal feature representation learning (Wang et al., 2018; Xue et al., 2021) and image captioning models (Zhou et al., 2020). Additionally, some approaches address the problem by classifying features related to writing style, sentiment, and user metadata (Przybyla, 2020). Propagation-based methods further enhance detection by allowing models to incorporate the context of social media activity, leveraging how users discuss and share information, which can serve as strong indicators of misinformation. Researchers typically employ neural networks such as GNN or True-LSTM for these models (Li et al., 2020; Ma et al., 2018). While promising, these approaches heavily rely on deep learning techniques, which often lack interpretability. Evidence-based approaches, on the other hand, aim to meticulously verify claims by integrating large amounts of external information. Early attempts at fact verification primarily focused on textual information like LIAR (Wang, 2017). Recently, an increasing number of multimodal datasets are being developed, drawing from fact-checking websites and social platforms with fact-checking articles, similar images, and user comments being collected and compiled as evidence datasets (Yao et al., 2023; Hu et al., 2023).

LLMs and MLLMs for Fact-Checking There has been growing interest in applying LLMs to fact-checking tasks. Hu et al. (2024a) demonstrated that while LLMs can offer multi-perspective analyses of fake news, their performance lags behind fine-tuned BERT models (Devlin et al., 2019). They further introduced the ARG method, selectively extracting insights from LLM responses, though their work did not incorporate evidence-based approaches. Wang et al. (2023) proposed an end-to-end solution for assessing the factuality of long texts by breaking documents into atomic check-worthy statements and verifying them through online evidence retrieval. Xuan et al. (2024) extended the retrieval-based approach to the multimodal domain, improving GPT-4V's misinformation detection with multi-round retrieval and topic

	Fauxtography	LatestSnopes	MR^2
Source	Snopes & Reuters	Snopes	Twitter
Begin	Jan 1998	Oct 2023	-
End	Jan 2019	Aug 2024	-
True	591	250	209
False	634	250	201
Total	1,225	500	410

Table 1: Statistics about the datasets in our evaluation. We use the English data from MR^2.

filtering. Fine-tuning approaches have also been explored to enhance fact-checking capabilities in open-source MLLMs. By fine-tuning instructions, these models can quickly adapt to specific fact-checking tasks. For example, Qi et al. (2024) developed Sniffer, which leverages the Google Entity Detection API[1] to identify visual entities in images and uses tailored prompts for GPT-4 to explain inconsistencies between image entities and textual claims. Wang et al. (2024) introduced MMIDR, which gathers image-centric evidence such as image captions, OCR data, and Google Reverse Image search results. Using this information, combined with ground truth data, they employed ChatGPT (gpt-3.5-turbo) to explain the rationale behind the claim labels. These efforts demonstrate the utility of fine-tuning open-source models by leveraging proprietary LLM generations.

3 Study Design

3.1 Evaluation Task

In this paper, our objective is to explore the capability of MLLMs in supporting fact-checking multimodal claims and the primary task is to detect the multimodal misinformation. Given an image-text pair (v, t) as a multimodal claim, the model is prompted to infer the veracity of the claim based on both its internal knowledge and external information e retrieved during the process. In addition to "True" and "False," the model is also allowed to output "Not Enough Information (NEI)" to express uncertainty in its assessment.

3.2 Approaches

Direct Prompting In this approach, the model is prompted to generate predictions for multimodal claims directly, without relying on external evidence or providing a reasoning process.

Chain of Thought (CoT) prompting has been demonstrated to significantly improve model performance in complex reasoning tasks (Wei et al., 2022). By generating a step-by-step reasoning process, the model can better leverage its internal knowledge and logical capabilities to address the challenge. Their intermediate reasoning steps offer greater transparency, allowing users to understand how the model arrives at its final prediction.

Retrieval Augmented Generation (RAG) has emerged as a powerful method for fact-checking by combining external knowledge sources with generative language models. It addresses the model's knowledge limitations by retrieving relevant online evidence to provide context for multimodal information. Like the CoT approach, we prompt the model to generate its reasoning process before making the final prediction.

To ensure a fair comparison between different models in the RAG approach, all models use the same evidence data. We employ both text-based and image-based retrieval methods to gather relevant evidence. For textual claims, we use the *DuckDuckGo Search API*[2] to retrieve the most relevant documents. For images, we leverage the *Google Reverse Image Search API*[3] to discover the most related documents. Additionally, we implement a filter to exclude evidence from fact-checking articles, as their clear conclusions might lead models to directly output the correct answer without sufficient reasoning. We parse and clean online documents with Trafilatura (Barbaresi, 2021), and use the date and text for evidence.

3.3 Datasets

To thoroughly evaluate the performance of MLLMs in fact-checking, we conduct experiments on three distinct datasets. ***Fauxtography*** is a multimodal dataset primarily derived from fact-checking articles on Snopes, along with images and headlines from Reuters news. The original dataset is binary, with labels indicating "True" or "False" claims. ***MR^2*** is a rumor detection dataset collected from social networking platforms, which includes an additional "Unverified" label. To mitigate issues related to outdated information, we filter out data that was unverified at the time of dataset construction from the test set, as it may have been resolved afterward. To mitigate data leakage and better eval-

[1] https://cloud.google.com/natural-language/docs/analyzing-entities

[2] https://pypi.org/project/duckduckgo-search/

[3] https://cloud.google.com/vision/docs/detecting-web

Model	Method	Fauxtography				LatestSnopes				MR^2			
		Prec.	Rec.	F1	NEI	Prec.	Rec.	F1	NEI	Prec.	Rec.	F1	NEI
InternVL2-4B	Direct	56.8	63.3	59.9	0.2	51.6	72.4	60.3	1.7	56.8	69.2	62.4	1.9
	CoT	60.7	69.1	64.6	0.3	53.2	86.5	65.9	2.3	58.5	82.1	68.3	2.4
	RAG	83.5	79.7	81.8	0.3	83.1	82.7	82.9	2.2	76.6	69.8	73.0	1.3
InternVL2-8B	Direct	63.7	54.8	58.9	7.0	51.2	77.6	65.6	10.7	60.6	71.2	65.5	17.3
	CoT	66.9	51.7	58.3	8.6	54.5	70.5	61.5	12.8	63.1	55.2	58.9	20.6
	RAG	86.0	80.0	82.9	2.0	79.5	89.7	84.3	1.5	75.9	67.7	71.6	7.7
InternVL2-26B	Direct	84.5	50.4	63.1	28.0	60.1	57.7	58.9	22.3	79.2	36.2	49.7	26.0
	CoT	87.1	40.7	55.5	28.9	63.2	55.0	58.9	32.3	87.5	34.8	49.8	27.6
	RAG	**94.1**	70.6	80.7	12.1	**94.7**	73.0	82.4	13.9	93.0	54.3	68.6	20.2
LLaVA-1.6-7B	Direct	72.2	56.2	63.2	33.4	56.6	40.1	46.9	46.2	89.2	33.7	48.9	50.6
	CoT	72.8	49.8	59.2	37.4	57.0	42.5	48.7	40.7	90.2	28.6	43.4	56.8
	RAG	80.7	72.6	76.4	12.9	84.7	75.6	79.9	14.8	72.8	68.7	70.7	21.2
LLaVA-1.6-13B	Direct	62.7	52.3	57.0	2.3	53.8	73.3	62.0	2.4	57.5	73.2	64.4	1.2
	CoT	65.7	58.3	61.8	5.3	57.1	74.4	64.6	5.8	58.4	69.4	63.4	4.6
	RAG	55.6	**92.5**	68.4	5.0	54.0	**90.5**	67.7	5.4	54.3	92.5	68.4	3.7
LLaVA-1.6-34B	Direct	79.5	54.2	64.5	21.4	59.2	52.7	55.8	24.4	72.8	40.7	52.2	26.5
	CoT	80.1	54.7	65.0	21.8	60.3	53.3	56.6	24.2	73.0	40.2	51.8	28.7
	RAG	89.2	71.3	79.2	7.6	83.2	76.7	79.8	11.3	87.0	61.2	71.8	16.3
GPT-4o	Direct	83.3	70.4	76.3	21.8	69.6	33.0	44.8	32.3	71.4	36.7	48.5	46.8
	CoT	90.8	77.6	83.7	19.4	71.1	64.0	67.4	27.3	**93.4**	42.8	58.7	40.7
	RAG	93.3	86.2	**89.6**	7.4	92.6	82.5	**87.3**	8.9	88.7	69.6	**78.0**	16.7

Table 2: Comparison of performance metrics (%) across various MLLMs, including GPT-4o, multiple size variants of InternVL2, and LLaVA-1.6. The approaches evaluated include Direct Prompting (Direct), Chain-of-Thought (CoT), and Retrieval-Augmented Generation (RAG). **Bold** and underlined numbers denote the best and the second best values in Precision (Prec.), Recall (Rec.) and F1-Score (F1).

uate the use of evidence, we have also collected the latest data from Snopes, ***LatestSnopes***, starting in October 2023, which corresponds to the knowledge cut-off date of GPT-4o[4]. We follow the construction methodology of *Fauxtography*. Table 1 shows the statistics of these datasets. Figure 3 in the Appendix presents the corresponding word clouds.

3.4 Experimental Setup

MLLMs We use OpenAI's API to obtain responses from GPT-4o and conduct experiments with open-source MLLMs including three variants of InternVL2 (4B, 8B, and 26B, Chen et al. 2023) and LLaVA-1.6 (7B and 13B and 34B, Liu et al. 2023b,a). It is important to note that the backbone LLMs of these multimodal models, despite being part of the same series, are developed by different institutions. Further details can be found in Table 6 in the Appendix. We empirically set the parameters for GPT-4o to $temperature = 0.01$ and $max_tokens = 300$ and do not investigate the impact of temperature and max_tokens on performance. We use the default settings for the open-source model without exploring the effects of temperature and max_tokens on performance. The prompts used for the different approaches are provided in Section A of the Appendix.

Metrics We evaluate performance using Precision, Recall, F1-Score, and NEI. NEI represents the proportion of instances where the model generates this response. For Precision, Recall, and F1-Score, misinformation detection is treated as a binary classification problem, with any instance where the model outputs "Not Enough Information (NEI)" considered an incorrect prediction.

4 Experimental Results and Analysis

4.1 Comprehensive Performance

Table 2 presents a comprehensive performance comparison of different models using various methods, and across multiple datasets. The values in RAG correspond to the use of a pair of evidence, consisting of the first pieces of evidence returned by text retrieval and reverse image search, respectively. Upon scrutinizing the results, we have the following key findings:

Data contamination allows GPT-4o to outperform other models on *Fauxtography* without

[4]https://otterly.ai/blog/knowledge-cutoff/

retrieved evidence. Open-source models perform close to random guessing without evidence. However, on the *Fauxtography* dataset, GPT-4o achieves significantly higher F1-Scores, outperforming other models even when evidence is present. Through string matching, we find that it flags 36 claims as "fact-checked," suggesting potential data leakage. This contamination is mitigated on the *LatestSnopes* and MR^2 datasets, where the higher NEI rate in the absence of evidence results in a drop in F1-Scores. Another reason for the high NEI rate on MR^2 is that it primarily consists of user posts from social platforms, many of which are semantically ambiguous. For example, a claim like "OMG I can't believe this!! Dad ran away!!" requires contextual information from external sources to fully grasp the claim's intent.

All models benefit from retrieval augmentation with relevant evidence. Notably, the integration of RAG boosts model performance across all datasets. GPT-4o retains its lead, showing a marked reduction in NEI responses and consistent improvements in Precision, Recall, and F1-Scores. The performance of open-source models also improves substantially. For example InternVL2-4B achieves F1-Scores of 81.8%, 82.9%, and 73.0% on *Fauxtography*, *LatestSnopes*, and MR^2, respectively, narrowing the gap with GPT-4o.

The performance of open-source models does not consistently correlate with model size. For InternVL models, NEI scores increase across all three datasets as the model size gets larger. Additionally, Precision also improves with larger model sizes across different methods. However, for LLaVA-1.6, we observe that the smallest model (LLaVA-v1.6-7B) tends to indicate uncertainty more frequently. LLaVA-1.6-13B demonstrates lower NEI scores compared to the 34B model, suggesting that a model's tendency to express uncertainty and its related performance are not solely determined by size, but may also be influenced by factors such as the various backbone LLMs and the fine-tuning methods. Furthermore, LLaVA-1.6-13B drops in Precision when evidence is available, while Recall increases.

4.2 Capabilities and Limitation Analysis

In this section, we explore the model's fine-grained capabilities in fact-checking, including detecting fake images and utilizing external evidence.

Model	RealandFakeFace				LatestSnopes			
	Prec.	Rec.	F1	NEI	Prec.	Rec.	F1	NEI
InternVL2-26B	**100.0**	8.5	15.7	0.0	77.8	23.9	36.6	0.0
LLaVA-1.6-34B	82.0	20.5	32.8	0.0	56.4	**86.3**	**68.2**	0.0
GPT-4o	92.5	**43.5**	**59.2**	26.5	**80.3**	45.6	58.2	36.3

Table 3: Performance (%) of various models on fake image detection.

4.2.1 Capability to Detect Fake Image

We observe that models' rationale in the CoT approach includes a detailed analysis of image authenticity, which occasionally contributes to more accurate predictions. To evaluate their ability to detect fake images, we assess them on two datasets without providing textual claims or retrieved evidence. The first dataset is RealandFakeFace (Nam et al., 2019), which contains expert-generated, high-quality photoshopped face images. We sample 200 real and 200 fake images for evaluation. The second dataset is sourced from *LatestSnopes*, where Snopes assigns a "Fake" rating to images that have been either digitally manipulated or AI-generated. This dataset includes 125 real and 125 fake images, covering a wide range of image content. We test InternVL2-26B, LLaVA-1.6-34B, and GPT-4o. The results in Table 3 show that while GPT-4o achieves high precision, its recall is low. LLaVA-1.6-34B exhibits substantial variation in Precision and Recall across the two datasets. **Overall, none of the three models can be considered reliable for detecting fake images.**

4.2.2 Capability to Utilize External Evidence

We examine the impact of the number of evidence pairs on model performance, as illustrated in Figure 1. The analysis is based on the results of InternVL models, as they demonstrated more consistent performance trends related to model size. We find that increasing the number of evidence pairs improves the accuracy of the InternVL models, with a larger improvement compared to GPT-4o. However, after the number of evidence pairs reaches 3, the gap between the InternVL models and GPT-4o no longer decreases. This suggests that the model reaches a bottleneck in multi-evidence retrieval scenarios. We further analyze the model's reasoning process and identify several limitations:

Ineffective or inadequate reasoning based on evidence We empirically find that the rationales generated by the InternVL model are generally shorter than those of GPT-4o. The detailed information are

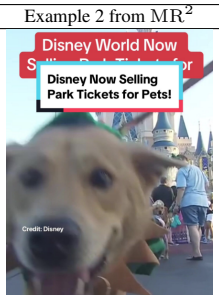
	Example 1 from LatestSnopes	Example 2 from MR^2
Multimodal Claim	On the final full day of his administration, the U.S. President Donald Trump awarded Dr. Anthony Fauci with a presidential commendation for his 'exceptional efforts on Operation Warp Speed,' the administration's project to develop a COVID-19 vaccine.	 Who is excited to bring their pets to Disney World?!
Ground Truth	True	False
Retrieved Document	Today, President Donald J. Trump awarded Presidential Commendations to the below individuals in recognition of their exceptional efforts on Operation Warp Speed: · · · Anthony Fauci Adam Boehler, Brad Smith, David Norquist, Chuong Huynh, Janssen Kimberly Taylor, David Simon, Rebecca Kurna, Jonathan Seals, Merck Daniel Wolfe, Christopher Houchens Robert · · ·	While the parks are for humans only (other than service animals), many locations around Disney World property are actually pet-friendly! If you can't bear the idea of leaving your furry best friend at home, you can always bring them with you on your next Disney trip. The Best Friends Pet Hotel helps to create magical experiences for your furry, four-legged friends (and more).
InternVL2-4B	The provided documents do not mention any awarding of a presidential commendation to Dr. Anthony Fauci on the final full day of Donald Trump's administration. While it is mentioned that Dr. Fauci was involved in Operation Warp Speed, there is no specific mention of a commendation being awarded to him. Therefore, the claim is false based on the information provided. Prediction: False	The visual of a dog looking directly at the camera suggests that there is excitement about bringing pets to Disney World. The text in the image further emphasizes that there are pet-friendly locations and services available at Disney World, such as the Best Friends Pet Hotel, pet-friendly hotels, and overnight boarding and daycare options. · · · Prediction: True
GPT-4o	Given the document provided, it lists Dr. Anthony Fauci as one of the individuals who were awarded Presidential Commendations for their "exceptional efforts on Operation Warp Speed." Operation Warp Speed was the administration's project to develop a COVID-19 vaccine. Since Dr. Anthony Fauci is mentioned in the provided document, and the image shows Dr. Fauci alongside then-U.S. President Donald Trump, it supports the claim that Dr. Fauci was indeed recognized by President Trump for his work on Operation Warp Speed. Prediction: True	The image shows a caption that claims "Disney Now Selling Park Tickets for Pets!", with a picture of a dog in front of Disney World. Given the information from the document that Disney's parks are for humans only aside from service animals, the claim made in the image contradicts the information in the document. Given that the document provides credible details about the actual pet accommodations at Disney World, the claim in the image is likely false and not a genuine policy change by Disney World. Prediction: False
InternVL2-4B + KT (*Ours*)	The document explicitly states that Dr. Anthony Fauci was awarded a Presidential Commendation by President Donald J. Trump for his exceptional efforts on Operation Warp Speed, which was the project to develop a COVID-19 vaccine. The image shows two men in front of a wall with a seal that includes the word "THE" and the silhouette of the White House, which suggests a governmental or official setting. This context is consistent with the information provided in the document, indicating that the claim is true. Prediction: True	The image depicts a dog in a costume, implying that pets are allowed at Disney World. However, the provided document indicates that Disney World parks are exclusively for humans. Therefore, the image may be conveying misleading information. Prediction: False.

Table 4: Examples of multimodal claims, retrieved documents, GPT-4 responses, and the changes observed in the open-source model InternVL2-4B before and after fine-tuning. Blue text indicates relevant information from the retrieved documents, yellow text highlights incorrect reasoning by the model based on the retrieved information, while red text marks correct reasoning in comparison.

shown in Figure 4 in the Appendix. Additionally, in many cases, the model fails to effectively leverage the retrieved evidence. For instance, in the first example from Table 4, although the retrieved documents contain relevant information supporting the claim, the model still produces incorrect reasoning. In the second example, the rationale generated by the InternVL model shows that its decision-making process relies solely on image analysis, completely ignoring the provided evidence, which also contributes to the shorter explanation length.

Biased focus For instance, in the claim, "A photo being shared online in May 2024 shows a pub in Lincolnshire selling a beer named 'Osama Bin Lager'" the InternVL model responds: "the documents do not provide any information about the specific month or year when the photo was shared." However, the date of the sharing is not the primary aspect of the fact-checking; the critical issue lies in verifying the existence of the pub in the claim. Despite relevant evidence confirming the pub's existence, InternVL incorrectly predicts the claim. Such errors are rarely observed in GPT-4o.

In summary, open-source models show obvious deficiencies compared to GPT-4o in their ability to effectively leverage evidence for reasoning.

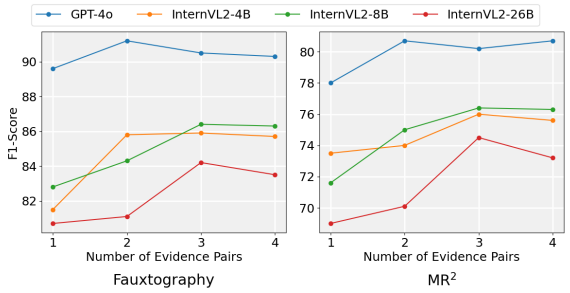


Figure 1: Impact of the number of evidence pairs on different models and datasets. Although an increased number of samples contributes to performance, the gap between open-source models and GPT-4 cannot be bridged merely by adding more evidence.

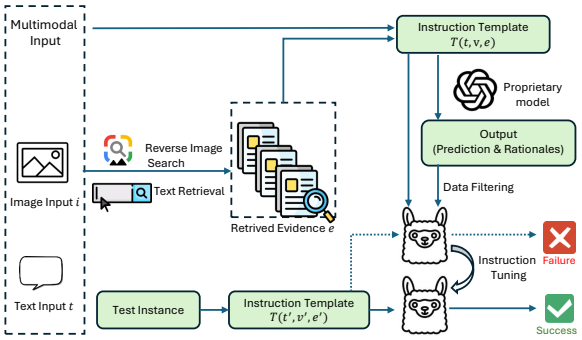


Figure 2: Architecture of the proposed framework. For multimodal inputs, we first collect different pieces of evidences. Next, we use prompt templates to gather veracity predictions and rationales from GPT-4o for each piece of multimodal information. Incorrect predictions are then filtered out, and finally, we fine-tune open-source models through instruction tuning.

5 Methodology

To address these limitations, a straightforward approach to improve performance is to construct instruction datasets and fine-tune the models. Given the high cost of manual annotation, we implement a *knowledge transfer* (KT) approach , leveraging GPT-4o as the teacher model by collecting its correct rationales. The open-source models are optimized to learn from this leading proprietary model, enhancing their capabilities in evidence-based fact-checking tasks. The architecture is illustrated in Figure 2. For each multimodal input (t, v), we collect 10 evidence pairs through reverse image search and text retrieval, excluding fact-checking websites as sources. To further enrich the evidence pool, we supplement it with articles from fact-checking websites. We then instantiate the instruction template with the multimodal information and evidence to generate the full prompt $T(t, v, e)$, which is fed into the proprietary GPT-4o model to obtain the corresponding generation. We filter out examples where the retrieved documents are not mentioned in the rationale. Ultimately, we obtain 11,962 prompt-response pairs, comprising 5,491 instances predicted as True, 5,568 predicted as False, and 903 predicted as NEI (Not Enough Information). Using the collected data and input prompts, we construct instruction datasets to fine-tune the models, ensuring that the same instruction template is used during both training and testing.

We fine-tune the InternVL2-4B and LLaVA-1.6-7B models using the collected data and official training scripts, applying LoRA for parameter-efficient tuning. Both models are trained for 10 epochs with the AdamW optimizer (Loshchilov and Hutter, 2019), using a batch size of 4. The learning rates are set at 4e-5 for InternVL2-4B and 2e-5 for LLaVA-1.6-7B. More details are in Section C in the Appendix.

Baselines We compare our method with the original model and existing instruction fine-tuning approaches. A VisualBERT-based binary classifier (Li et al., 2019) is trained with two variants: one using claims only and another combining claims with evidence pairs. We implement Sniffer (Qi et al., 2024) and MMIDR (Wang et al., 2024) based on the *Fauxtography* dataset as well. Sniffer utilizes the Google Entity Detection API to identify entities within images and prompts GPT-4o to analyze inconsistencies between the visual entities and the corresponding textual claims. Predictions flagged as inconsistent are classified as False. MMIDR constructs instruction datasets by retrieving relevant images and extracting captions using Baidu's PaddleOCR (Du et al., 2020) and BLIP-2 OPT-2.7B (Li et al., 2023a), then prompting GPT-4o to generate rationales based on both ground truth and the retrieved documents. It is worth noting that in both Sniffer and MMIDR, the teacher models do not directly access images.

5.1 Results

The final results are presented in Table 5, with MLLM results obtained using the RAG approach, following the same setup as in Table 2. Compared to the original models, the models fine-tuned with our method (KT) show varying degrees of improvement across different metrics. Through knowledge transfer, the F1-Score of the InternVL2-4B model on the LatestSnopes dataset improved from

Model	LatestSnopes				MR^2			
	Precision	Recall	F1-Score	NEI	Precision	Recall	F1-Score	NEI
VisualBERT	82.4	77.1	79.6	-	74.8	73.5	74.1	-
+ Evidence	87.4	83.1	85.2	-	79.2	**76.7**	77.9	-
InternVL2-4B	83.1	82.7	82.9	2.2	76.6	69.8	73.0	1.3
+ Sniffer	80.5	76.6	78.5	3.2	72.7	69.8	71.1	1.7
+ MMIDR	84.7	**86.6**	85.6	2.0	78.6	75.5	77.0	0.9
+ KT (*Ours*)	90.1	83.1	86.5	5.9	84.3	72.7	77.5	8.9
- NEI cases	87.0	84.2	85.6	2.3	80.9	73.1	76.8	2.2
LLaVA-1.6-7B	84.7	75.6	79.9	14.8	72.8	68.7	70.7	21.2
+ Sniffer	76.2	79.1	78.0	2.2	74.8	72.1	73.5	5.5
+ MMIDR	86.7	78.4	82.3	3.5	77.8	72.6	75.1	6.2
+ KT (*Ours*)	89.2	80.7	84.7	10.4	82.9	70.1	75.9	12.3
- NEI cases	87.8	81.1	84.3	3.0	80.2	71.4	75.5	5.6
GPT-4o	**92.6**	82.5	**87.3**	8.9	**88.7**	69.6	**78.0**	16.7

Table 5: Comparison (%) of instruct-tuning-based approaches and baseline models. MLLMs generate responses based on the same retrieved evidence.

82.9% to 86.5%, while the LLaVA-1.6-7B model increased from 79.9% to 84.7%. Similar improvements were observed on the MR^2 dataset. This demonstrates the effectiveness of our approach, which outperforms other baselines across different models. We also compare the results when excluding NEI cases, meaning the training data only includes predictions labeled as True or False, similar to the setup used in Sniffer and MMIDR. The results show that filtering out these samples lowers the NEI scores and impairs the model's predictive ability, with F1-Scores decreasing to varying degrees. VisualBERT performs better when evidence is available; however, it cannot generate rationales, limiting its explainability. The Sniffer method does not improve the model's ability to predict veracity. For InternVL2-4B, after fine-tuning, the F1-Scores drop from 82.9% and 73.0% to 78.5% and 71.1%, respectively. This indicates that simply checking for consistency between images and text is insufficient for detecting misinformation. The MMIDR method also shows improvements across different datasets and models; however, it falls short of our approach. Manual inspection of the model's reasoning revealed inaccuracies, likely due to training data where the teacher was required to reason based on insufficient or irrelevant evidence.

We apply string matching to quantify the proportion of instances where the model references "document" or "evidence." Our findings demonstrate that instruction fine-tuning significantly improves this rate. For InternVL2-4B, the proportion increases from 76.4% and 70.2% to 90.0% and 85.7% on the *LatestSnopes* and MR^2 2 datasets, respectively. Similarly, for LLaVA-1.6-7B, the numbers rise from 70.2% and 70.5% to 86.4% and 82.8%.This trend can also be observed in Table 4, in the first example, the model correctly states, "The document explicitly states that Dr. Anthony Fauci was awarded a Presidential Commendation by President Donald J. Trump for his exceptional efforts on Operation Warp Speed." In the second example, the model also references evidence, stating, "The provided document indicates that Disney World parks are exclusively for humans. Therefore, the image may be conveying misleading information."

6 Conclusion and Future Work

We conducted a comprehensive evaluation of existing MLLMs in fact-checking, including proprietary models such as GPT-4o open-source ones like InternVL2 and LLaVA-1.6. We found that even using the leading GPT-4o significantly decreases in the absence of external knowledge, highlighting the importance of retrieval. Moreover, open-source models improve their performance considerably through RAG, though a gap remains compared to proprietary models. Upon examining the rationales, we found that many critical pieces of evidence were not accurately used. To address this, we proposed a knowledge transfer framework that leverages instruction-based fine-tuning to learn GPT-4o's reasoning approach. This proved effective across multiple datasets.

In future work, we plan to evaluate more multimodal models, to use a wider variety of prompting techniques, as well as to explore the performance of MLLMs for multilingual claims.

Limitations

We acknowledge several limitations. (1) We did not thoroughly investigate the sensitivity of different models to various prompts, which may lead to an overestimation or underestimation of certain models' performance. (2) We did not conduct an in-depth analysis of NEI (Not Enough Information) responses, which could provide valuable insights for multi-round retrieval and optimizing search queries. (3) Although *LatestSnopes* is constructed from debunking articles published after GPT-4o's knowledge cut-off date, it is still possible that the events occurred earlier, which cannot entirely prevent data contamination.

Ethics and Broader Impact

Bias: We acknowledge that our evaluation dataset includes celebrities and social events, which might raise concerns regarding privacy and reputation; however, all data originates from publicly available sources or websites.

Misuse Potential: Our assessment of multimodal models, particularly GPT-4o, in fact-checking should not be construed as an endorsement of absolute reliance on these models, especially in contexts involving critical information or data.

Environmental Impact: The training of large models is energy-intensive and could contribute to climate change. Fortunately, our experiments primarily focus on inference rather than training these large models, thereby mitigating potential environmental impacts.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal automated fact-checking: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5430–5448, Singapore. Association for Computational Linguistics.

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022a. GERE: Generative evidence retrieval for fact verification. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'22, pages 2184–2189, Madrid, Spain.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022b. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Tsun-Hin Cheung and Kin-Man Lam. 2023. FactLLaMA: Optimizing instruction-following language models with external knowledge for automated fact-checking. *ArXiv preprint*, abs/2309.00240.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. 2020. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024a. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of Thirty-Eighth AAAI Conference on Artificial Intelligence*, AAAI 2024, pages 22105–22113, Vancouver, Canada. AAAI Press.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. 2024b. Towards understanding factual knowledge of large language models. In *The Twelfth International Conference on Learning Representations*, ICLR'2024, Vienna, Austria.

Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S. Yu. 2023. Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2901–2912, New York, NY, USA. Association for Computing Machinery.

Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros. 2018. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, ECCV 2018, Munich, Germany.

Mamadou Keita, Wassim Hamidouche, Hessen Bougueffa Eutamene, Abdelmalik Taleb-Ahmed, and Abdenour Hadid. 2024. Fidavl: Fake image detection and attribution using vision-language model. *arXiv preprint arXiv:2409.03109*.

Jiawen Li, Yudianto Sujana, and Hung-Yu Kao. 2020. Exploiting microblog conversation structures to detect rumors. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5420–5429, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Zizhong Li, Haopeng Zhang, and Jiawei Zhang. 2023b. A revisit of fake news dataset with augmented fact-checking by chatgpt. *ArXiv preprint*, abs/2312.11870.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *ArXiv preprint*, abs/2310.03744.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023*, NeurIPS'2023, New Orleans, LA, USA.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6817, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on Twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989, Melbourne, Australia. Association for Computational Linguistics.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Seonghyeon Nam, Seoung Wug Oh, Jae Yeon Kang, Chang Ha Shin, Younghyun Jo, Young Hwi Kim, Kyungmin Kim, Minho Shim, Sungho Lee, Yunji Kim, Suho Han, Gunhee Nam, Dasol Lee, Subin Jeon, In Cho, Woongoh Cho, Sejong Yang, Dongyoung Kim, Hyolim Kang, Sukjun Hwang, and Seon Joo Kim. 2019. Real and fake face detection, version 1. `https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection`. Retrieved [Date Retrieved].

Piotr Przybyla. 2020. Capturing the style of fake news. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):490–497.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, CVPR'2024, Seattle, Washington, United States. Computer Vision Foundation / IEEE.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

Longzheng Wang, Xiaohan Xu, Lei Zhang, Jiarui Lu, Yongxiu Xu, Hongbo Xu, and Chuang Zhang. 2024. Mmidr: Teaching large language model to interpret multimodal misinformation via knowledge distillation. *arXiv preprint arXiv:2403.14171*.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 849–857, New York, NY, USA. Association for Computing Machinery.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, et al. 2023. Factcheck-gpt: End-to-end fine-grained document-level fact-checking and correction of llm output. *arXiv preprint arXiv:2311.09000*.

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

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

Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. *Inf. Process. Manage.*, 58(5).

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. 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, page 2733–2743, New York, NY, USA. Association for Computing Machinery.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *ArXiv preprint*, abs/2303.18223.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. : Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on knowledge discovery and data mining*, pages 354–367. Springer.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. Fact-checking meets fauxtography: Verifying claims about images. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2099–2108, Hong Kong, China. Association for Computational Linguistics.

A Prompts

Prompt for direct prompting

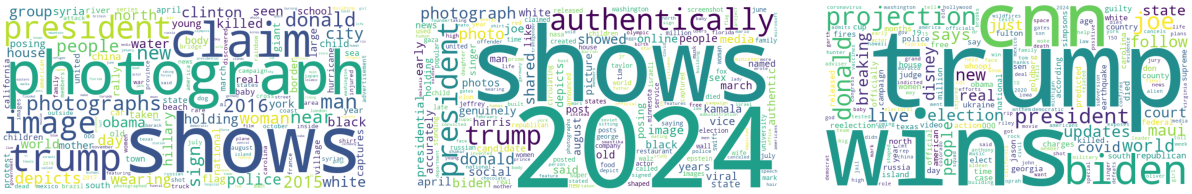
```
You are a professional fact-checking
assistant and are required to use the
specified format to provide your response.

Given an IMAGE and the following CLAIM:
CLAIM.

Please predict the truth of the multimodal
claim as either True, False, or NEI (Not
Enough Information).

Prediction: [True/False/NEI]
```

Prompt for Chain-of-Thought

```
You are a professional fact-checking
assistant and are required to use the
specified format to provide your response.

Given an IMAGE and the following CLAIM:
CLAIM.

Please predict the truth of the multimodal
claim as either True, False, or NEI (Not
Enough Information).

Prediction: [True/False/NEI]

Let's think step by step,
```

Prompt for Retrieval Augmented Generation

```
You are a professional fact-checking
assistant and are required to use the
```



(a) Fauxtography (b) LatestSnopes (c) MR^2

Figure 3: Word clouds visually represent the frequency and importance of key topics across three datasets.

```
specified format to provide your response.

You are first given several retrieved documents.
Documents: DOCUMENTS

Given an IMAGE and the following CLAIM: CLAIM.

Please based on the documents to predict the truth of the multimodal claim as either True, False, or NEI (Not Enough Information).

Prediction: [True/False/NEI]

Let's think step by step.
```

Prompt for Fake Image Detection

```
You are an advanced AI model trained to assess images for authenticity. Your task is to analyze the given image and determine whether it is real or fake. Provide a clear and concise response with your prediction and an explanation supporting your decision.

Please output your response in the following format:

Prediction: [Real or Fake or NEI(Not Enough Information)]

Let's think step by step.
```

B More Statistics and Resutls

Topic Distribution of Datasets. As shown in Figure 3, we apply TF-IDF analysis to highlight the most significant terms in each dataset, and present the topics of various datasets using word clouds to visualize the most prominent terms, providing an intuitive overview of the dataset's content.

Open-source MLLMs. Table 6 presents the architectures of the open-source MLLMs used in our study. For the LLaVA-1.6 (Liu et al., 2024), the vision encoder remains the same while the backbone LLMs vary. For the InternVL2 (Chen et al., 2023), the 4B variant utilizes a language model from Phi-3 (Abdin et al., 2024), whereas the other two models use language models trained by the institution itself (Cai et al., 2024). For further details, please refer to the original works.

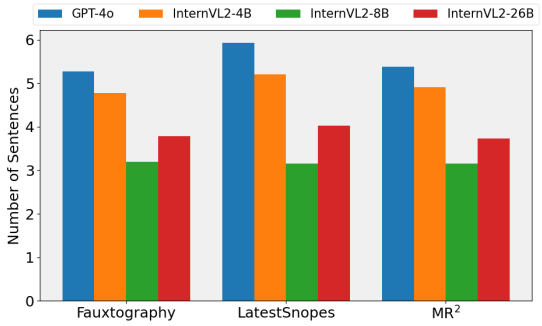


Figure 4: Number of sentences in different MLLM's rationales.

MLLMs	Vision Part	Language Part
InternVL2-4B	InternViT-300M-448px	Phi-3-mini-128-instruct
InternVL2-8B	InternViT-300M-448px	Internlm2_5-7b-chat
InternVL2-26B	InternViT-6B-448px-V1-5	Internlm2-chat-20b
LLaVA-1.6-7B	CLIP-ViT-L-336px	Mistral-7B-Instruct-v0.2
LLaVA-1.6-13B	CLIP-ViT-L-336px	Vicuna-13b-v1.5
LLaVA-1.6-34B	CLIP-ViT-L-336px	Nous-Hermes-2-Yi-34B

Table 6: Vision and language components of the open-source models used in our experiments.

Number of sentences in the MLLMs' rationales. Figure 4 shows the number of sentences in the rationales generated by different GPT-4o models and InternVL2 models. GPT-4o consistently produces more than five sentences in its rationales across all three datasets, exceeding the output of other models. Additionally, the sentence length of InternVL2-4B surpasses that of the two larger open-source models.

C Training Details

All our experiments were conducted on NVIDIA A100 GPUs with 80GB of memory. Full instruction tuning on our dataset takes approximately 14 hours.