

Multimodal Large Language Models to Support Real-World Fact-Checking

Jiahui Geng, Yova Kementchedjhieva, Preslav Nakov, Iryna Gurevych

Mohamed bin Zayed University of Artificial Intelligence

{jiahui.geng, yova.kementchedjhieva,preslav.nakov,iryna.gurevych}@mbzuai.ac.ae

Abstract

Misinformation poses a real-world threat, especially when combined with misleading images. Multimodal large language models (MLLMs), which combine image comprehension with the rich knowledge and explanatory capability of language models, have become tools for humans to process vast information. However, their capacity and limitations as multimodal tools in assisting with fact-checking remain understudied. Here is aim to bridge this gap. In particular, we propose a framework for systematically assessing the capacity of current multimodal models to facilitate real-world fact-checking. Our methodology is evidence-free, leveraging only these models' intrinsic knowledge and reasoning capabilities. By designing prompts that extract models' predictions, explanations, and confidence levels, we conduct a detailed analysis of the model's accuracy, bias, and other key factors. We empirically find that (1) GPT-4V exhibits surprising performance across various datasets, with an accuracy rate exceeding 80%, and is capable of providing impressive explanations, and (2) even with the aid of prompt ensembles and in-context learning, open-source models significantly lag in performance. However, they show potential in remembering checked claims and reasoning out manipulated images. We also summarize the failure reasons, which contributes to strategies for future improvements. Our study offers insights into leveraging MLLMs to combat multimodal misinformation.

1 Introduction

Misinformation is a significant challenge on the internet, especially with regard to multimodal claims, which combine text, images, videos, and

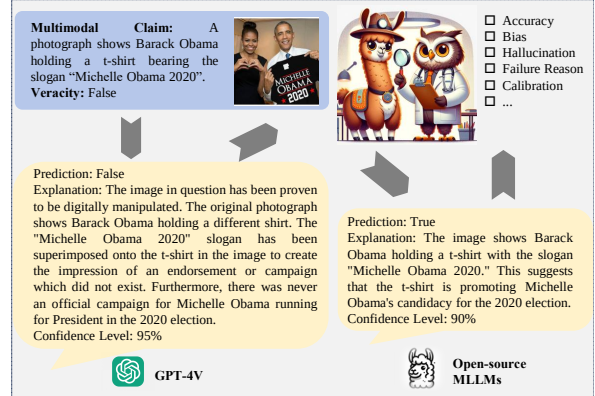


Figure 1: Illustration of our proposed framework to evaluate the capability of MLLMs as fact-checkers. Initially, we collect their responses to multimodal claims, encompassing predictions, explanations, and confidence levels. We then assess their performance across dimensions, including accuracy, bias, and their failure reasons.

other media types (Akhtar et al., 2023). The visual component in these cases could be manipulated or used out-of-context (OOO) to make a false claim (Huh et al., 2018; Luo et al., 2021; Aneja et al., 2023; Yao et al., 2023). In such cases, fact-checkers and the tools they employ need to be able to handle multiple modalities.

Large language models (LLMs) store extensive information, which exceeds what any individual human can know and is accessible in a far more human-friendly way than any search engine can offer (Zhao et al., 2023; Hu et al., 2024). As such, they can be a powerful tool in the hands of fact-checkers, who often seek extra factual knowledge to verify a claim (Nakov et al., 2021). The application of LLMs to fact-checking has been extensively studied in a text-only setting, wherein a model receives as input both a textual claim, with or without a set of evidence documents, and

needs to reason to determine the veracity of the claim (Chen et al., 2022; Pan et al., 2023; Li et al., 2023a). LLMs have increasingly strong understanding and reasoning capabilities, so using them to sift through evidence documents when verifying a claim is proving very effective (Chen et al., 2022; Li et al., 2023a). Yet, this pipeline approach requires an extra step of evidence retrieval, which is computationally heavy and error-prone, while it overlooks the vast amount of knowledge stored in the parameters of LLMs (Hu et al., 2023). We hypothesize that multimodal large language models (MLLMs) trained on large amounts of data can serve as a sufficient substitute for the retrieval of evidence and perform fact-checking in an evidence-free fashion, i.e., relying solely on their parametric knowledge.

Despite their great promise, MLLMs have limitations as well, specifically with respect to factuality (Fu et al., 2023; Liu et al., 2023c). Models may produce hallucinations (Huang et al., 2023; Li et al., 2023b) and suffer from a lack of knowledge or exhibit biases (Hu et al., 2024; Wang et al., 2023). To understand MLLMs’ potential to support fact-checking real-world claims, we propose the evaluation framework illustrated in Figure 1. It involves querying models for predictions, explanations, and confidence scores on multimodal check-worthy claims. Based on the data obtained, we can further assess the models’ capabilities from various perspectives, addressing the following research questions:

- **RQ1 Performance Evaluation:** How good are MLLMs in identifying multimodal misinformation? Can they perform accurate reasoning?
- **RQ2 Improving Approaches:** How can we effectively boost the model’s fact-checking capabilities? Can the improvements be reflected in the reasoning, beyond accuracy metrics?
- **RQ3 Taxonomy of Failure Reasons:** What are the typical errors for MLLMs employed as multimodal fact-checkers?

Our investigation spans an array of vision-language models, including GPT-4V, MiniGPT-v2 (Chen et al., 2023) and LLaVA-1.5 (Liu et al., 2023b), which stand as leaders in this domain. We evaluate these models across three existing

multimodal fact-checking datasets: Fauxtography (Zlatkova et al., 2019), COSMOS (Aneja et al., 2023), MOCHeg (Yao et al., 2023), and one new dataset, which build from recent check-worthy claims to mitigate the risk of data contamination. We use prompt ensembles (PE) and in-context learning (ICL) to help improve the open-source MLLMs. We do not use fine-tuning because models are required to explain its reasoning process and provide confidence scores, while high-quality annotated data is scarce. Our study finds that MLLMs display a range of fact-checking capabilities. GPT-4V, in particular, excels in our tests, exhibiting high accuracy, useful explanations, and good calibration. The accuracy on various datasets generally reaches around 80%. MiniGPT-v2 lacks explanatory and uncertainty-reporting capabilities. LLaVAs can follow instructions, and both PE and ICL enhance their performance to varying degrees, with the latter providing greater improvement. On Fauxtography and COSMOS, ICL increases the Overall Accuracy from 52.3% to 62.3% and from 67.7% to 76.5%, respectively. The majority baseline for both datasets is 50%.

In summary, we systematically evaluate existing multimodal models to determine their capability to support real-world fact-checking. Our approach is more comprehensive than prior studies, which are limited to specific aspects such as face spoofing, forgery detection, or out-of-context detection (Shi et al., 2024; Qi et al., 2024). We formulate several research questions to evaluate these models based on accuracy, bias, reasoning capabilities, and errors across a variety of datasets. Our work fills a research gap, providing a comprehensive overview of the domain and shedding light on leveraging MLLMs to combat multimodal misinformation.

2 Related Work

LLMs for Text-Only Fact-Checking Caramanion (2023) compared the performance of models such as ChatGPT 3.5 and ChatGPT 4.0 in news fact-checking. FactLLaMA (Cheung and Lam, 2023) integrates external evidence into the instruct-tuning process to enhance the model’s ability to leverage evidence for predictions. Studies have also focused on leveraging the capabilities of LLMs to develop automated frameworks for decomposing claims, generating queries, and pre-

dicting based on gathered evidence, thereby augmenting the interpretability of the fact-checking process (Pan et al., 2023; Li et al., 2023a). However, these works often focused solely on prediction accuracy. Our work analyzes the model’s fact-checking capabilities from three perspectives: predictions, explanations, and confidence.

Multimodal Fact-Checking Multimodal misinformation primarily falls into two categories: one involves textual claims about manipulated content across different modalities, while the other pertains to out-of-context misinformation, featuring unaltered images, audio, or videos in misleading contexts (Luo et al., 2021; Aneja et al., 2023; Yao et al., 2023). Shi et al. (2024) proposed a benchmark to evaluate the performance of MLLMs in detecting face spoofing and forgery. Qi et al. (2024) introduced Sniffer, an MLLM designed for OOC detection and explanation, utilizing fine-tuning to boost the accuracy in identifying celebrities and clarifying inconsistencies between text and images. In our work, we posit that evidence is parametrically encoded within the models, enabling us to bypass the evidence retrieval phase and directly engage in predictions and explanations for the real-world claims.

3 Evaluation Framework

We propose an evaluation framework that includes datasets, prompts, and evaluation metrics to address the research questions.

3.1 Datasets

Due to constraints posed by the GPT-4V API, such as decreased query speeds after exceeding a daily limit, we sample data from various datasets and preprocess them to fit our evaluation framework. Additionally, we develop a new dataset to ensure the model has not encountered these data.

Fauxtography is a multimodal fact-checking dataset sourced from the websites Snopes and Reuters (Zlatkova et al., 2019). We select a random subset of 400 entries, ensuring a balanced composition of 200 true and 200 false instances.

COSMOS is a dataset built from Snopes and News Outlets, and tailored for OOC detection (Aneja et al., 2023). We randomly select 240 true and 240 false samples from the test split. Each image in this dataset comes with two captions, and

	Fauxtography	COSMOS	MOCHEG	Post-4V
True	200	240	267	81
False	200	240	237	105
Total	400	480	504	186

Table 1: Statistics of datasets in our evaluation.

we manually select one caption per image to ensure it aligns with our criteria.

MOCHEG is originally based on textual claims from Snopes and PolitiFact, with associated images serving as evidence (Yao et al., 2023). This implies that a single claim can correspond to multiple images, many of which are not explicitly required for the claim to be verified. We first identify multimodal claims through keywords such as *photograph*, *image*, etc. Subsequently, we manually select images that were explicitly requested for verification and filter out those with overlaying words like *fake*, *misleading*, or *miscaptioned*. This procedure results in a total of 504 data entries.

Post-4V We collect new data from Snopes, focusing on articles published after the release of GPT-4V (September 26, 2023). We apply the same filtering methodology used in MOCHEG, resulting in a dataset of 186 samples.

3.2 Evaluation Prompt

We simultaneously obtain the predictions, explanations, and confidence levels from MLLMs with the prompt below. This prompt is formulated following the example of prompts found in related fact-checking work (Pan et al., 2023; Min et al., 2023), while for acquiring explanations and confidence estimates we refer to Xiong et al. (2024). We collect their verbalized confidence as it is increasingly used for decision-making during human-machine collaboration (Geng et al., 2023).

Is it true that "**CLAIM**"? True or False? Use the following format to provide your answer:
Prediction: [True or False]
Explanation: [put your evidence and step-by-step reasoning here]
Confidence Level: [please show the percentage]

Note: The confidence level indicates the degree of certainty you have about your answer and is represented as a percentage. For instance, if your confidence level is 80%, it means you are 80% certain that your answer is correct and there is a 20%

chance that it may be incorrect.

3.3 Evaluation Metrics

Response Types Based on the three components (prediction, explanation, and confidence level) in the model’s response, we categorize them into four types. *Others* refer to instances where the model fails to respond according to our requirements, missing any component. *Uncertain* denotes cases where the model explicitly expresses uncertainty, with statements such as “cannot confirm”, “cannot verify”, or “cannot determine”, etc. The remaining samples are grouped as *True* or *False* according to the model’s prediction. Overall, the numbers of these four indicators reveal the model’s ability to follow instructions, express uncertainty, and inherent bias in predicting True and False.

Accuracy Metrics We design two accuracy metrics to reflect the model’s performance. One is *True&False Accuracy*, measuring the accuracy of samples that are solely classified as True or False. This is important because users usually do not rely on the model’s response when the model clearly expresses uncertainty or does not follow instructions. The other metric, named *Overall Accuracy*, calculates the proportion of responses that contain correct predictions across all samples. This facilitates comparison between different approaches.

4 Experimental Setups

We use OpenAI’s API to collect responses from GPT-4V (gpt-4-1106-preview). In addition, we experiment with open-source MLLMs, including LLaVA-1.5 (7b and 13b, Liu et al. 2023b,a), and MiniGPT (v2, Chen et al. 2023). All parameters are set to the default values, with *max_tokens*=300 for GPT-4V. These models are notably representative and competitive in a multimodal setup (Fu et al., 2023; Liu et al., 2023c). We further explore two approaches that do not require fine-tuning to enhancing the model’s performance:

Prompt Ensembles (PE) involve using a variety of prompts for the same task and aggregating model’s responses to produce more accurate and less biased predictions. We employ ChatGPT to generate 5 more semantically similar yet distinct prompts, as depicted in Figure 2. The part about explanations and confidence scores is identical to the original prompt. With these prompts we col-

lect six responses for each claim. Then, we conduct a vote to determine the final response type. If the two highest-scoring response types receive the same number of votes, we consider the model uncertain.

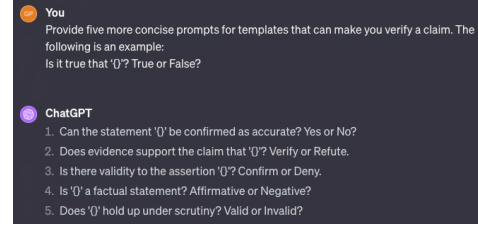


Figure 2: Prompts obtained from ChatGPT that are used in prompt ensembles experiments.

In-Context Learning (ICL) enables a model to learn tasks by directly observing examples within the prompt, bypassing the need for prior explicit training. Debunking articles, which are lengthy and often contain irrelevant descriptions, are not used here. Instead, we utilize responses from GPT-4V to claims not included in our test datasets as instructive examples. To overcome LLaVA’s limitation in handling multiple image inputs, we vertically stack different images with a 20-pixel separation. The relationship between each image and its corresponding claim is explicitly described, such as "For the first picture, claim: ...", as shown in the prompt below. We perform 1 and 2-shot learning, showing the model one and two examples, respectively, referred to as ICL-1 and ICL-2. We use four pairs of examples as demonstrations. For ICL-1, all examples are False, while for ICL-2, the first claim is always False, and the second is True. They are detailed illustrated in Figure 10 in Appendix.

Use the following format to answer whether the claim regarding the pictures is True or False:

SAME FORMAT AS IN THE ORIGINAL PROMPT.

For the first picture, claim: "**CLAIM1**"
Prediction: False
Explanation: XXX
Confidence Level: XXX

For the second picture, claim: "**CLAIM2**"
Prediction: True
Explanation: XXX
Confidence Level: XXX

For the third picture, claim: "**CLAIM**"

	Fauxtography						COSMOS						MOCHEG						Post-4V					
	T	F	U	O	T&F	All	T	F	U	O	T&F	All	T	F	U	O	T&F	All	T	F	U	O	T&F	All
GPT-4V	158	195	29	18	81.9	82.1	179	204	83	14	86.2	80.0	216	223	37	28	87.2	83.5	54	98	26	8	79.6	73.7
MiniGPT	-	-	-	-	-	55.5	-	-	-	-	-	62.1	-	-	-	-	-	54.3	-	-	-	-	-	55.9
LLaVA(7b)	337	6	1	56	53.4	46.3	449	7	4	20	52.0	50.0	409	6	0	89	54.2	44.8	157	1	3	25	42.0	37.1
LLaVA(13b)	286	98	3	13	54.4	52.3	331	134	13	2	69.7	67.7	363	112	28	1	61.5	58.1	116	65	2	3	56.4	55.9
LLaVA+PE	244	153	2	1	57.1	54.7	275	204	0	1	76.3	71.7	290	214	0	0	59.9	58.1	85	101	0	0	56.9	56.1
LLaVA+ICL-1	228	159	6	8	61.8	62.3	293	175	9	3	74.8	74.1	254	240	7	4	62.4	62.5	79	101	5	1	58.7	57.5
LLaVA+ICL-2	186	188	8	18	61.6	61.7	247	215	12	7	77.3	76.5	195	286	10	13	60.2	60.4	48	122	8	8	62.1	61.4
Majority	-	-	-	-	-	50.0	-	-	-	-	-	50.0	-	-	-	-	-	53.0	-	-	-	-	-	56.5

Table 2: Performance of various models and approaches. *T*: True, *F*: False, *U*: Uncertain, *O*: Others, *T&F*: True&False Accuracy, *All*: Overall Accuracy, *PE*: Prompt Ensembles, *ICL*: In-Context Learning. The majority-class accuracy is established in the last row.

5 Experimental Results

We comprehensively analyze the data we collected in Section 5.1. In Section 5.2, we provide a detailed analysis of the model explanations, focusing on their step-by-step reasoning, and discuss our taxonomy of failure reasons. Section 5.3 is dedicated to ablation studies.

5.1 Comprehensive Results

The comprehensive results are displayed in Table 2. The first four columns for each dataset represent the number of different types of responses, while the following are True&False and Overall Accuracy, respectively. We have highlighted the highest metrics and the highest ones among the open-source models. Hereafter, we omit the version numbers of the open-source models and refer to them simply as LLaVA(7b), LLaVA(13b), and MiniGPT. Both PE and ICL experiments are based on LLaVA(13b). LLaVA+PE involves voting based on 6 responses; if the counts of True and False are equal, then it is an uncertain case. LLaVA+ICL-1 and LLaVA+ICL-2 represent the average results across four sets of demonstrations, we have rounded the number of different responses. Detailed outcomes for PE and ICL are presented in Table 5 and 6 in Appendix.

Overall, GPT-4V demonstrates surprising accuracy, with both accuracy metrics exceeding 80% on the three public datasets. The accuracy decrease is not particularly significant on Post-4V, with the True&False Accuracy equals 79.6%. In addition, these values are mostly higher than Overall Accuracy, suggesting that GPT-4V exhibits higher precision when it responds with confidence. There are more False cases than True ones. The high number of *Others* is primarily due to GPT-4V

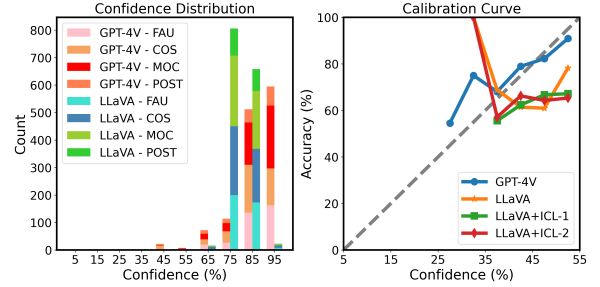


Figure 3: The left graph illustrates the confidence score distribution of GPT-4V and LLaVA(13b), and the right graph presents their calibration curves. *FAU*: Fauxtography, *COS*: COSMOS, *MOC*: MOCHEG, *POST*: Post-4V.

safety alignment. This results in frequent replies such as "I'm sorry, but I cannot assist with this request." or "This content may violate our content policy." Additionally, GPT-4V expresses uncertainty the most.

In contrast, MiniGPT cannot provide explanations and confidence, while LLaVA(7b) shows a strong bias, almost always responding True with this prompt. LLaVA(13b) shows improved accuracy across all datasets, better adherence to instructions, and more frequent expressions of uncertainty. We observe that the two approaches, PE and ICL, can enhance the model's performance to varying degrees. As to PE, there is an obvious improvement in the COSMOS dataset, with increases of 6.6 and 4.0 in two accuracy metrics, respectively. However, the enhancement on other datasets is very limited. ICL shows a greater impact. On the COSMOS dataset, LLaVA+ICL-1 brings increases of 7.6 and 9.0 in both metrics. On MOCHEG, the improvement is minimal, with gains of 0.9 and 4.4, respectively.

	Fauxtography & Veracity: False	COSMOS & Veracity: True	MOCHEG & Veracity: False	Post-4V & Veracity: True
Multimodal Claim	A photograph shows Barack Obama holding a t-shirt bearing the slogan "Michelle Obama 2020." 	A Red Cross burial worker shows a man how to put on protective gloves. 	A photograph captured Joe Biden posing menacingly with an elderly woman and a gun. 	A photograph authentically shows pioneering software engineer Margaret Hamilton standing next to the code she wrote by hand that took humanity to the moon in 1969. 
GPT-4V w. Image	Prediction: False Explanation: The image in question has been proven to be digitally manipulated. The original photograph shows Barack Obama holding a different shirt. The "Michelle Obama 2020" slogan has been superimposed onto the t-shirt in the image to create the impression of an endorsement or campaign which did not exist. Furthermore, there was never an official campaign for Michelle Obama running for President in the 2020 election. Confidence Level: 95%	Prediction: True Explanation: ... which is typically worn for activities requiring infection control measures such as handling hazardous materials or dealing with infectious diseases. The other man is being shown how to put on a blue protective glove. ..., and the presence of a crowd suggests a community-based educational or training event, which is in line with the activities routinely conducted by organizations such as the Red Cross in areas affected by epidemics or contagious diseases. Confidence Level: 90%	Prediction: False Explanation: The image presented appears to be a satirical or manipulated image. "Google image search: 'creepy Joe Biden'" implies that it might be a meme or an image made for humorous purposes. Joe Biden, the former Vice President and current President of the United States as of my last training cut-off in 2023, has been in the public eye for many years, and any such legitimate photograph would have garnered significant attention from reputable news sources. ... when they seem outlandish or meant for humor. Confidence Level: 95%	Prediction: True Explanation: The image in question features Margaret Hamilton, a pioneering computer scientist and software engineer who led the development of the onboard flight software for the Apollo missions to the Moon for NASA during the 1960s. The stack of books beside her is often described as the listings of the Apollo Guidance Computer (AGC) source code. ... The photograph has been widely attributed to represent her standing next to the code she and her team wrote for the Apollo program. Confidence Level: 95%
LLaVA w. Image	Prediction: True Explanation: The image shows Barack Obama holding a t-shirt with the slogan "Michelle Obama 2020." This suggests that the t-shirt is promoting Michelle Obama's candidacy for the 2020 election. Confidence Level: 90%	Prediction: True Explanation: The image shows a ... This suggests that the woman is a Red Cross burial worker, and the man is learning how to properly wear gloves for safety purposes. The presence of a crowd in the background further supports the idea that this is a public event or gathering where safety measures are being demonstrated. Confidence Level: 80%	Prediction: True Explanation: The image shows Joe Biden posing menacingly with an elderly woman and a gun. The woman is holding her tongue out, and the gun is pointed at her. This suggests that the photograph captures a threatening situation involving Joe Biden and the elderly woman. Confidence Level: 80%	Prediction: False Explanation: The image shows two women standing next to a large stack of books, not Margaret Hamilton standing next to the code she wrote by hand. There is no evidence in the image to support the claim that the code she wrote by hand took humanity to the moon in 1969. Confidence Level: 80%
LLaVA w/o. Image	Prediction: False Explanation: I do not have any evidence or information to support the claim that Barack Obama is holding a t-shirt with the slogan "Michelle Obama 2020." Additionally, it is unlikely that Barack Obama would be endorsing his wife's potential presidential campaign in such a manner. Confidence Level: 90%	Prediction: True Explanation: The image shows a man wearing gloves and a Red Cross burial worker demonstrating how to put on protective gloves. The presence of the Red Cross burial worker and the man wearing gloves suggests that the worker is teaching the man how to put on protective gloves. Confidence Level: 90%	Prediction: False Explanation: Joe Biden is a public figure and it is highly unlikely that he would pose menacingly with an elderly woman and a gun. Additionally, there is no evidence to support this claim. Confidence Level: 90%	Prediction: False Explanation: Margaret Hamilton is a pioneering software engineer who played a crucial role in the development of the Apollo 11 guidance computer. However, the statement "...," is not true. There is no photograph that shows Margaret Hamilton standing next to the code she wrote by hand. Confidence Level: 90%
LLaVA w. Image 1 example	Prediction: True Explanation: The image shows ... This suggests that the photo is a campaign advertisement or promotional material for Michelle Obama's potential candidacy in the 2020 presidential election. The presence of Barack Obama in the image further supports this claim, as he was the former President of the United States and is likely to be involved in his wife's political campaign. Confidence Level: 90%	Prediction: True Explanation: In the image, there is a person wearing a hazmat suit and a face mask, which indicates that they are taking precautions to protect themselves from potential hazards. ... The presence of a Red Cross burial worker also supports the idea that the scene involves a situation that requires specialized protection and guidance. Confidence Level: 80%	Prediction: False Explanation: ... the presence of a gun in the image is likely a manipulation or a misinterpretation of the scene. Joe Biden is a politician and former Vice President of the United States, and it is highly unlikely that he would pose with a gun in a menacing manner. The image appears to be a genuine photograph of Joe Biden and an elderly woman, but the gun and menacing pose are added elements that are not part of the original scene. Confidence Level: 90%	Prediction: True Explanation: The image features a woman standing next to a stack of books, which are the code she wrote by hand. The books are arranged in a way that resembles the Apollo 11 guidance computer code. The woman is Margaret Hamilton, a pioneering software engineer who played a crucial role in the development of the Apollo 11 guidance computer. The image is an authentic representation of her contribution to the Apollo 11 mission. Confidence Level: 85%

Figure 4: Sampled fact-checking responses from different models and approaches. The first row shows the claim source and its veracity. The second row includes multimodal claims, and the subsequent four rows feature responses from GPT-4V, LLaVA(13b), LLaVA(13b) without image input, and LLaVA+ICL-1 (using the first demonstration), respectively. Purple text indicates hallucinations by the model when no images are present; red text shows outdated knowledge, and green text displays the model’s analysis of image manipulation.

Calibration The left part of Figure 3 displays the confidence distributions of GPT-4V and LLaVA(13b), breaking down the data into 10% confidence intervals. For instance, a value of 95 corresponds to confidence levels within the (90, 100] bracket. GPT-4V’s confidence scores are largely clustered in the (80, 100] interval, whereas LLaVA(13b)’s scores are more frequently found in the (70, 90] range. The right calibration curves

reveal that both GPT-4V and LLaVA exhibit a positive correlation between confidence levels and accuracy, with GPT-4V’s calibration closely mirroring the ideal curve (illustrated by the dashed line), indicating well-calibrated confidence scores. Conversely, LLaVA(13b)’s curve suggests a propensity for overconfidence. We find that ICL does not result in better calibration of LLaVA’s verbalized confidence as their curves are very close.

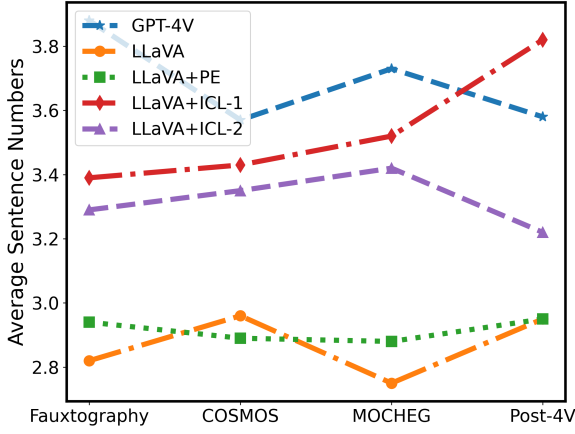


Figure 5: Average number of sentences in explanations across different models and settings. GPT-4V generates the longest explanations except on Post-4V. With one example, ICL-1 significantly increases the average explanation length.

5.2 Explanation-based Analysis

5.2.1 Case Studies

Figure 4 displays totally 4 cases with the original multimodal claims and corresponding responses. Rows labeled "GPT-4V w. Image," "LLaVA w. Image," and "LLaVA w. Image 1 example" correspond to the results of GPT-4V, LLaVA(13b), and LLaVA+ICL-1, respectively. It can be observed that GPT-4V’s explanation contains a richer set of facts. For instance, it explicitly states, "there was never an official campaign for Michelle Obama in the 2020 election," in the 1st case, and in the 4th case, it provides detailed background information about Margaret Hamilton, including her role in "leading the development of onboard flight software for the Apollo mission." Moreover, GPT-4V adeptly incorporates details from the images, such as referencing a "Google image search: Creepy Joe Biden" in the 3rd image.

However, LLaVA(13b) lacks the capability to counteract manipulated images, as evidenced in the 1st and 3rd cases. Additionally, LLaVA(13b)’s explanations are typically brief, often limited to a description of the image followed by a conclusion about its veracity. It shows a lack of effective reasoning, rendering the explanations less comprehensive than GPT-4V. We find ICL enhances LLaVA(13b)’s fact-checking capabilities. For instance, the model accurately analyzes the 3rd image, with "a genuine photograph of Joe Biden and an elderly woman, but the gun and menacing pose are added elements that are not part of the original

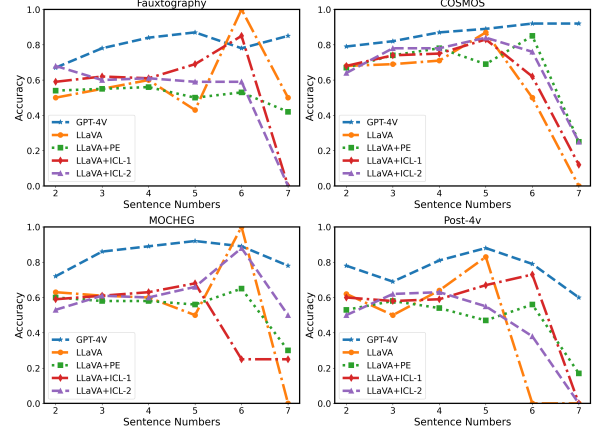


Figure 6: The relationship between response’s explanation length and the response accuracy. Generally, accuracy increases with the length of the sentences, peaking at around 5 or 6 sentences, after which it significantly declines.

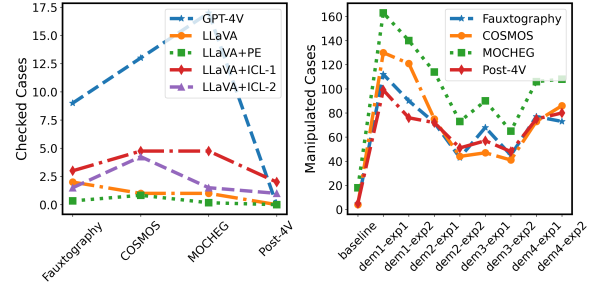


Figure 7: The left graph presents the number of checked cases across various datasets and approaches, with the numbers for LLaVA+PE, LLaVA+ICL-1, and LLaVA+ICL-2 representing average counts. The graph on the right illustrates the comparison of the quantity of manipulated cases identified when using and not using the examples. The baseline is LLaVA(13b).

scene." In the 4th case, LLaVA+ICL-1 provides a more detailed description of Margaret Hamilton’s background and states that "the image is an authentic representation of her contribution." However, it still fails in the 1st case with ICL.

5.2.2 Analysis of Explanation Length

For convenience, we refer to the number of sentences in the explanation part of a response as the explanation length. The average length of model explanations with different approaches is shown in Figure 5. We use NLTK to segment the sentences. GPT-4V provides the highest number of sentences in explanations across various datasets, with a minimum of 3.57 sentences in COSMOS

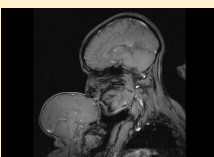
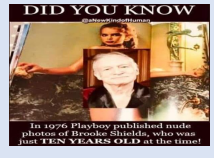
Lack of Knowledge. It's common knowledge that octopuses have eight tentacles, the rarity of an octopus with 96 tentacles may confound the model.	Similar Images. The image depict a scene from the movie "Tropic Thunder," but the model mistakes it for the "Platoon." Both movies share the Vietnam War as background.
Claim: A photograph shows an octopus with 96 tentacles. Veracity: True  An octopus with 96 tentacles found in Japan in 1998	Claim: A photograph shows a group of underappreciated Vietnam veterans. Veracity: False 
Distortion due to Perspective and Lens. This example demonstrates that GPT-4V does not understand that a telephoto lens can make the foreground appear closer to the background than it actually is.	Incomplete Verification. The model's prediction overlooks the fact that an MRI cannot capture chemical reactions.
Claim: A photograph of a fire at Notre Dame Cathedral in Paris shows the Eiffel Tower directly in the background. Veracity: True 	Claim: An MR image captures a mother's kiss producing a 'chemical reaction' that released a burst of oxytocin in her baby's brain. Veracity: True 
Blurry or Incomplete Image. The model cannot discern that the scene depicted is related to posture training based on the image's content..	Ambiguous Claims. Playboy can refer to either the name of the publisher or the magazine.
Claim: A photograph shows a pin in the collar of a Chinese officer's uniform, a training method to correct posture. Veracity: True 	Claim: Playboy published nude photographs of 10-year-old actress Brooke Shields. Veracity: True 

Figure 8: Taxonomy of failure reasons based on GPT-4V’s responses. Each block includes the response from GPT-4V, along with explanation and reason for its failure.

and a maximum of 3.88 in Fauxtography. The value of LLaVA+PE represents the average length of explanations under different prompts. ICL prompts LLaVA to generate more sentences; however, providing two examples (LLaVA+ICL-2) does not yield longer explanations than providing just one (LLaVA+ICL-1), the average length even decreases.

We further analyze the relationship between the model’s prediction accuracy and the explanation length, as depicted in Figure 6, where we display explanation lengths ranging from 2 to 7, encom-

passing the most samples generated by various models. Typically, different models achieve peak accuracy when the explanation length is 5 or 6 sentences. Specifically, on COSMOS, the accuracy of the different models grows steadily as the explanation length increases from 2 to 5. Across the other three datasets, the left half of the curves all show varying degrees of ascent despite noticeable fluctuations. GPT-4V achieves its highest accuracy at an explanation length of 5 sentences on them.

	Fauxtography						COSMOS						MOCHEG						Post-4V					
	T	F	U	O	T&F	All	T	F	U	O	T&F	All	T	F	U	O	T&F	All	T	F	U	O	T&F	All
GPT-4V	105	202	32	61	80.1	71.0	111	253	50	66	64.0	57.5	153	265	38	48	73.9	68.9	35	68	10	73	74.8	67.7
LLaVA(7b)	398	0	2	0	50.3	50.0	474	0	4	2	50.2	49.8	502	0	2	0	53.0	52.8	181	0	4	1	43.1	43.0
LLaVA(13b)	167	214	17	2	62.5	62.0	263	195	20	2	76.6	75.8	168	316	20	0	62.0	61.5	40	124	22	0	60.4	60.8

Table 3: Performance of various models without image input. Red numbers highlight those have increased compared to those with images. Text-only claims decrease the accuracy of GPT-4V but increase the accuracy of LLaVA (13b).

5.2.3 Reasons for Predicting False

To better understand the model’s fact-checking process, particularly why it labels multimodal claims as false, we explore two types of cases: checked cases, the model confirms that statements have been verified by organizations; manipulated cases, it indicates that images have been altered.

Checked Cases We count the occurrences of phrases such as "been checked" and "been debunked" across different settings, with the results displayed in left graph in Figure 7. GPT-4V mentions these keywords most frequently across three public datasets, significantly more than others. Additionally, ICL prompts LLaVA to mention them more often. Notably, our ICL examples do not include checked cases. However, on Post-4V, several claims are considered fact-checked by LLaVA, but not by GPT-4V. The accuracy of GPT-4V and LLaVA+ICL on checked cases is 86.5% and 88.8%, respectively. The average True&False Accuracy for those that are not checked cases is only 84.4% and 65.7%, respectively.

Manipulated Cases We notice that only the examples in the 3rd demonstration do not include manipulated cases. Therefore, we compare the impact of demonstrations on the number of manipulated cases, as shown in the right of Figure 7. Despite not accessing manipulated cases as example, the number significantly rises in demo3-1-exp1 and demo3-exp2, although these figures remain lower compared to other demonstrations. Thus, reasoning from the perspective of image manipulation is also not an explicit learning process. The accuracy of LLaVA is 67.6% for manipulated cases and 65.4% for non-manipulated cases, respectively. This demonstrates LLaVA’s potential to analyze veracity from the perspective of image manipulation.

5.2.4 Taxonomy of Failure Reasons

Upon examining GPT-4V’s explanations, we identify six primary categories of failure reasons. *Lack of Knowledge*, especially obscure knowledge that contradicts intuition, easily leads the model to assess a statement’s truthfulness inaccurately. *Similar Images* refer to instances where the model associates with another image featuring a similar scene. In such cases, the model incorrectly extracts knowledge, leading to inconsistent content with the presented image. *Distortion Due to Perspective and Lens*, we find that GPT-4 cannot correctly understand images that are distorted due to the shooting angle or lens.. *Incomplete Verification* refers to the cases where the model does not fully consider the atomic claims within a multimodal claim. For example, when a claim contains both correct and incorrect subclaims, or when the textual claim is accurate but the accompanying image is misleading. *Blurry or Incomplete Image* can lead to an incorrect prediction due to the model’s uncertainty about what it sees. *Ambiguity Claims* involve semantic uncertainties, where a term could reference multiple entities. We show representative examples and explanations for each failure reason in Figure 8. Our taxonomy can lead to targeted optimization strategies. For instance, we could include more images with distortions or blurry or incomplete images from image scaling and cropping into the instruction-tuning dataset.

5.3 Ablation Studies

Impact of Images To understand the impact of images on fact-checking, we conduct a comparative experiment where we ask GPT-4V and LLaVA (13b) to re-evaluate the veracity of textual claims without providing images. As shown in Table 3, the results are marked in red for values that exceed those in Table 2. GPT-4V experiences a decrease in accuracy across all datasets, while surprisingly, LLaVA (13b) shows higher accuracy compared to

	Fauxtography						COSMOS						MOCHEG						Post-4V					
	T	F	U	O	T&F	All	T	F	U	O	T&F	All	T	F	U	O	T&F	All	T	F	U	O	T&F	All
GPT-4V	158	158	72	12	80.7	63.8	162	190	109	18	79.8	70.0	234	204	54	33	83.7	77.8	51	83	38	14	80.1	72.6
LLaVA(7b)	237	71	7	49	54.1	50.5	311	101	14	54	66.0	60.1	341	89	8	66	56.0	50.8	125	37	3	21	51.2	49.5
LLaVA(13b)	207	143	7	42	51.7	52.0	228	201	22	28	69.9	68.5	235	212	7	50	61.8	59.5	65	105	2	14	61.8	60.2

Table 4: Performance of various models when reasoning first. Red numbers highlight those that have increased compared to those predicting first. Reasoning first significantly increases the expression of uncertainty in GPT-4V.

when images are included. To understand this, we further look into their explanations. As shown in Figure 4, in the 1st and 3rd cases, the model can make correct predictions based on internal knowledge without images, stating "it is unlikely that Barack Obama would be endorsing his wife's ..." and "Joe Biden is a public figure and it is highly unlikely that he would pose menacingly ...". However, the model is misled when manipulated images are presented. The GPT-4V's responses without image input are in Figure 9 in Appendix.

Reasoning First In the experiments above, we prompt the model to first output a prediction, followed by step-by-step reasoning. We conduct a comparative experiment to assess the impact of reversing this order—starting with reasoning before prediction—on accuracy. The results are presented in Table 4, where values exceeding those in Table 2 are highlighted in red. We observe that starting with reasoning significantly increases the instances of GPT-4V expressing uncertainty. For example, the number of uncertain cases rises from 29 to 72 on Fauxtography and 10 to 38 on Post-4V. This increase directly leads to an obvious decline in Overall Accuracy, although True&False Accuracy remains relatively high. LLaVA(7b) still predicts most claims as True. LLaVA(13b) shows varying accuracy improvements across the last three datasets, though the changes are not substantial. We also note an increase in instances where the model fails to follow instructions, which may partially influence the accuracy metrics.

6 Limitations

We acknowledge the following limitations in our work. First, due to constraints in manpower and computational resources, we did not test the complete Fauxtography and COSMOS datasets. The limited number of samples may impede an accurate assessment of model characteristics. Second, we did not quantify the different capabilities

within fact-checking, such as the ability of different models to recognize AI-generated images, identify faces and scenes, and retrieve contextual knowledge. Moreover, we did not investigate how fine-tuning with domain-specific data could potentially improve model performance. Future research will aim to address these gaps and provide a more comprehensive evaluation of the models.

7 Conclusion and Future Work

We investigated the capabilities of MLLMs to fact-check real-world claims, relying solely on their parametric knowledge and reasoning capabilities without external references. We proposed an evaluation framework, designed various experiments to address the research questions. Our results indicated that state-of-the-art MLLMs, such as GPT-4V, have the potential to assist professional fact-checkers: they can enhance their efficiency by providing reference predictions, valuable clues, and explanatory insights, together with confidence. However, open-source models like LLaVA fall behind. They can be misled by manipulated images and may generate hallucinations when no image is present. We empirically demonstrate that prompt ensembles (PE) and in-context learning (ICL) are able to improve model's accuracy in detecting misinformation, while ICL is more efficient.

In future work, we plan to investigate how to enhance the fact-checking capabilities of models, focusing not only on improving accuracy but also on strengthening robustness and augmenting the knowledge reasoning and the verification abilities required for fact-checking. A potential approach is to use GPT-4V to enhance the capabilities of smaller open-source models through knowledge distillation. We also plan to explore how MLLMs can better support fact-checking when connected to external knowledge.

References

- Mubashara Akhtar, Michael Schlichtkrull, Zhi-jiang 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.
- Shivangi Aneja, Chris Bregler, and Matthias Nießner. 2023. [COSMOS: Catching out-of-context image misuse using self-supervised learning](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37 of AAAI 2023, pages 14084–14092, Washington, DC, USA.
- Kevin Matthe Caramancion. 2023. [News verifiers showdown: a comparative performance evaluation of chatgpt 3.5, chatgpt 4.0, bing ai, and bard in news fact-checking](#). *arXiv preprint arXiv:2306.17176*.
- Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022. [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.
- Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunsyong Xiong, and Mohamed Elhoseiny. 2023. [MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning](#). *ArXiv preprint*, abs/2310.09478.
- 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.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. [MME: A comprehensive evaluation benchmark for multimodal large language models](#). *ArXiv preprint*, abs/2306.13394.
- Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koepl, Preslav Nakov, and Iryna Gurevych. 2023. [A survey of language model confidence estimation and calibration](#). *ArXiv preprint*, abs/2311.08298.
- Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. 2024. [Towards understanding factual knowledge of large language models](#). In *The Twelfth International Conference on Learning Representations*, ICLR’2024, Vienna, Austria.
- Xuming Hu, Zhaochen Hong, Zhijiang Guo, Lijie Wen, and Philip Yu. 2023. [Read it twice: Towards faithfully interpretable fact verification by revisiting evidence](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR’23, pages 2319–2323, Taipei, Taiwan.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *ArXiv preprint*, abs/2311.05232.
- 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.
- Miaoran Li, Baolin Peng, and Zhu Zhang. 2023a. [Self-Checker: Plug-and-play modules for fact-checking with large language models](#). *ArXiv preprint*, abs/2305.14623.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. [Evaluating object hallucination in large vision-language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.
- 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, 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.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. [MMBench: Is your multi-modal model an all-around player?](#) *ArXiv preprint*, abs/2307.06281.
- 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.
- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FActScore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. 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.
- Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. [Fact-checking complex claims with program-guided reasoning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.
- Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. [Sniffer: Multimodal large language model for explainable out-of-context misinformation detection](#). *arXiv preprint arXiv:2403.03170*.
- Yichen Shi, Yuhao Gao, Yingxin Lai, Hongyang Wang, Jun Feng, Lei He, Jun Wan, Changsheng Chen, Zitong Yu, and Xiaochun Cao. 2024. [SHIELD: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models](#). *arXiv preprint arXiv:2402.04178*.
- Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023. [Survey on factuality in large language models: Knowledge, retrieval and domain-specificity](#). *ArXiv preprint*, abs/2310.07521.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. [Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs](#). In *Proceeding of The Twelfth International Conference on Learning Representations, ICLR’2024*, Vienna, Austria.
- 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.
- 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)*,

A More Experimental Results

Figure 9 displays additional responses, where "w. Image" indicates multimodal claims as input and "w/o. Image" denotes text-only claims as input. The last two rows show the results for LLaVA+ICL-1 and LLaVA+ICL-2, respectively. As shown in Figure 9, GPT-4V's explanations demonstrate its awareness of missing images, employing phrases such as "there is no verified public record" or "without any provided context." This indicates that GPT-4V actively retrieves knowledge to verify the presence or context of images. Compared with no demonstrations, the results reveal that ICL enhances the model's reasoning capabilities, incorporating more detailed analysis of images and discussions about the motivations behind their creation.

Table 5 shows the results with different prompts in prompt ensembles (PE). The accuracy of the GPT-4V model is quite consistent across various prompts. In contrast, the LLaVA model is very sensitive to the prompts, with some prompts showing significantly higher accuracy than others on the same dataset. However, no single prompt consistently performs best across all datasets. Table 6 displays the impact of different demonstrations in ICL on the LLaVA (13b) model. Demonstrations clearly enhance the model's accuracy. However, the benefit of the second example is limited, and it does not show a significant improvement compared to the ICL-I cases.

	Fauxtography & Veracity: False	COSMOS & Veracity: True	MOCHEG & Veracity: False	Post-4V & Veracity: True
Multimodal Claim	A photograph shows Barack Obama holding a t-shirt bearing the slogan "Michelle Obama 2020." 	A Red Cross burial worker shows a man how to put on protective gloves. 	A photograph captured Joe Biden posing menacingly with an elderly woman and a gun. 	A photograph authentically shows pioneering software engineer Margaret Hamilton standing next to the code she wrote by hand that took humanity to the moon in 1969. 
GPT-4V w. Image	Prediction: False Explanation: The image in question has been proven to be digitally manipulated. The original photograph shows Barack Obama holding a different shirt. The "Michelle Obama 2020" slogan has been superimposed onto the t-shirt in the image to create the impression of an endorsement or campaign which did not exist. Furthermore, there was never an official campaign for Michelle Obama running for President in the 2020 election. Confidence Level: 95%	Prediction: True Explanation: ... which is typically worn for activities requiring infection control measures such as handling hazardous materials or dealing with infectious diseases. The other man is being shown how to put on a blue protective glove. ..., and the presence of a crowd suggests a community-based educational or training event, which is in line with the activities routinely conducted by organizations such as the Red Cross in areas affected by epidemics or contagious diseases. Confidence Level: 90%	Prediction: False Explanation: The image presented appears to be a satirical or manipulated image. "Google image search: 'creepy Joe Biden'" implies that it might be a meme or an image made for humorous purposes. Joe Biden, the former Vice President and current President of the United States as of my last training cut-off in 2023, has been in the public eye for many years, and any such legitimate photograph would have garnered significant attention from reputable news sources. ... when they seem outlandish or meant for humor. Confidence Level: 95%	Prediction: True Explanation: The image in question features Margaret Hamilton, a pioneering computer scientist and software engineer who led the development of the onboard flight software for the Apollo missions to the Moon for NASA during the 1960s. The stack of books beside her is often described as the listings of the Apollo Guidance Computer (AGC) source code. ... The photograph has been widely attributed to represent her standing next to the code she and her team wrote for the Apollo program. Confidence Level: 95%
GPT-4V w/o. Image	Prediction: False Explanation: There is no verified public record or credible source to validate that such a photograph of Barack Obama holding a t-shirt with the slogan "Michelle Obama 2020" exists. A quick search on the internet and major news outlets does not produce any result matching the description. Barack Obama has publicly commented that Michelle Obama has no intention to run for president. Confidence Level: 95%	Prediction: False Explanation: Without any provided context, specific details, or source material to evaluate and verify this claim, I cannot confidently affirm that a Red Cross burial worker shows a man how to put on protective gloves. The Red Cross does engage in health and safety education, but the assertion's specifics can't be confirmed without further information. Confidence Level: 50%	Prediction: False Explanation: Without additional context or a specific photograph to refer to, it's ambiguous. However, ... does not reveal any credible evidence or reliable report featuring a photograph of Joe Biden posing menacingly with an elderly woman and a gun ... Furthermore, Joe Biden as President supports stricter gun control measures, making such an event even more implausible. Confidence Level: 95%	Prediction: True Explanation: Margaret Hamilton is indeed a pioneering software engineer who developed software for NASA's Apollo Project which was instrumental in the success of the 1969 moon landing. A famous photograph shows her standing next to the stack of code she and her team wrote for Apollo's onboard flight software. Hamilton is often credited as being instrumental in the creation of software engineering as a field. Confidence Level: 95%
LLaVA w. Image	Prediction: True Explanation: The image shows Barack Obama holding a t-shirt with the slogan "Michelle Obama 2020." This suggests that the t-shirt is promoting Michelle Obama's candidacy for the 2020 election. Confidence Level: 90%	Prediction: True Explanation: The image shows a ... This suggests that the woman is a Red Cross burial worker, and the man is learning how to properly wear gloves for safety purposes. The presence of a crowd in the background further supports the idea that this is a public event or gathering where safety measures are being demonstrated. Confidence Level: 80%	Prediction: True Explanation: The image shows Joe Biden posing menacingly with an elderly woman and a gun. The woman is holding her tongue out, and the gun is pointed at her. This suggests that the photograph captures a threatening situation involving Joe Biden and the elderly woman. Confidence Level: 80%	Prediction: False Explanation: The image shows two women standing next to a large stack of books, not Margaret Hamilton standing next to the code she wrote by hand. There is no evidence in the image to support the claim that the code she wrote by hand took humanity to the moon in 1969. Confidence Level: 80%
LLaVA w/o. Image	Prediction: False Explanation: I do not have any evidence or information to support the claim that Barack Obama is holding a t-shirt with the slogan "Michelle Obama 2020." Additionally, it is unlikely that Barack Obama would be endorsing his wife's potential presidential campaign in such a manner. Confidence Level: 90%	Prediction: True Explanation: The image shows a man wearing gloves and a Red Cross burial worker demonstrating how to put on protective gloves. The presence of the Red Cross burial worker and the man wearing gloves suggests that the worker is teaching the man how to put on protective gloves. Confidence Level: 90%	Prediction: False Explanation: Joe Biden is a public figure and it is highly unlikely that he would pose menacingly with an elderly woman and a gun. Additionally, there is no evidence to support this claim. Confidence Level: 90%	Prediction: False Explanation: Margaret Hamilton is a pioneering software engineer who played a crucial role in the development of the Apollo 11 guidance computer. However, the statement "... is not true. There is no photograph that shows Margaret Hamilton standing next to the code she wrote by hand." Confidence Level: 90%
LLaVA w. Image 1 example	Prediction: True Explanation: The image shows ... This suggests that the photo is a campaign advertisement or promotional material for Michelle Obama's potential candidacy in the 2020 presidential election. The presence of Barack Obama in the image further supports this claim, as he was the former President of the United States and is likely to be involved in his wife's political campaign. Confidence Level: 90%	Prediction: True Explanation: In the image, there is a person wearing a hazmat suit and a face mask, which indicates that they are taking precautions to protect themselves from potential hazards. ... The presence of a Red Cross burial worker also supports the idea that the scene involves a situation that requires specialized protection and guidance. Confidence Level: 80%	Prediction: False Explanation: ... the presence of a gun in the image is likely a manipulation or a misinterpretation of the scene. Joe Biden is a politician and former Vice President of the United States, and it is highly unlikely that he would pose with a gun in a menacing manner. The image appears to be a genuine photograph of Joe Biden and an elderly woman, but the gun and menacing pose are added elements that are not part of the original scene. Confidence Level: 90%	Prediction: True Explanation: The image features a woman standing next to a stack of books, which are the code she wrote by hand. The books are arranged in a way that resembles the Apollo 11 guidance computer code. The woman is Margaret Hamilton, a pioneering software engineer who played a crucial role in the development of the Apollo 11 guidance computer. The image is an authentic representation of her contribution to the Apollo 11 mission. Confidence Level: 85%
LLaVA w. Image 2 examples	Prediction: True Explanation: The image shows ... This suggests that the photo was taken during a campaign event or a political rally, where the former president was promoting his wife's candidacy for the 2020 presidential election. The presence of the t-shirt and the context of the photo indicate that the claim is accurate. Confidence Level: 90%	Prediction: True Explanation: The image ... he is following proper hygiene and safety protocols while handling a dead body. The presence of a Red Cross burial worker suggests that this scene is related to a humanitarian effort, such as a disaster relief or a burial team providing assistance to those in need. The man in the mask and gloves is likely... ensure both his safety and the safety of others during the handling process. Confidence Level: 80%	Prediction: False Explanation: ... The photo has been digitally altered to create a false and misleading representation of the situation. Joe Biden is a former U.S. Vice President and has never been associated with any violent or threatening behavior towards an elderly woman or anyone else. ... The image is likely created to spread misinformation and discredit Joe Biden or to provoke a reaction from viewers. Confidence Level: 95%	Prediction: True Explanation: ... Margaret Hamilton is known for her significant contributions to the Apollo 11 mission, where she led the development of the onboard flight software that guided the spacecraft to the moon. ... The photograph in question is widely recognized as a historical image that captures her role in the Apollo 11 mission. The presence of the code she wrote by hand in the photograph further supports the authenticity of the image and her involvement in the mission. Confidence Level: 80%

Figure 9: More detailed responses from GPT-4V and LLaVA(13b) across various experimental setups. Purple text indicates hallucinations by the model when no images are present, red text shows outdated knowledge, and green text displays the model’s analysis of image manipulation.

Demo1		Textual Claim: A photograph shows a man surfing with a whale. Prediction: False Explanation: The image you've provided appears to be digitally manipulated. While surfing and whales are both associated with the ocean, the scale and proximity of the whale to the surfer in this photograph are not consistent with typical whale behavior or the physics of ocean waves. Whales are generally not found surfing in waves, and their size compared to the surfer seems disproportionate. Additionally, the water disturbance caused by a whale of that size would likely be much more significant than what is depicted here. The lack of any news coverage or scientific documentation of such an extraordinary event further suggests that this image is not an accurate representation of reality. Confidence Level: 95%
		Textual Claim: A photograph shows Morgan Freeman on the children's show 'The Electric Company'. Prediction: True Explanation: The man in the photograph is indeed Morgan Freeman. He began his career with appearances in several children's educational programs, including "The Electric Company," which aired from 1971 to 1977. In "The Electric Company," Freeman played several characters, including the iconic Easy Reader, which is likely the character depicted in this photograph, as indicated by his attire and the overall style of the set. This aligns with the known history of Freeman's roles and the time period in which the show was popular. Confidence Level: 90%
Demo2		Textual Claim: A photograph shows rays of lights in the shape of a cross during the August 2017 eclipse. Prediction: False Explanation: The claim that a photograph shows rays of light arranged in a perfect cross shape during the August 2017 eclipse is not supported by astronomical events or the physics of how light behaves during an eclipse. During a total solar eclipse, phenomena such as Bailey's beads or the diamond ring effect may occur, which can result in spectacular visual displays as the moon covers or uncovers the sun. However, a cross shape made of light rays is not a typical occurrence and could instead be a sign of digital manipulation or an artifact of the photography process rather than a natural event. A natural cross shape in the sky during an eclipse would be an extraordinary claim requiring substantial evidence, which is not typically reported or documented by reliable sources in astronomy. Confidence Level: 95%
		Textual Claim: A photograph captured in mid-June 2021 showed the Robertson Draw Fire burning a mountainside outside of Red Lodge, Montana. Prediction: True Explanation: The image provided shows a large fire on a mountainside in the background with a clear foreground of a street with buildings, vehicles, and an American flag, characteristic of a small town setting. Given the description of the photo matching the image and knowing that wildfires in Montana around mid-June of 2021, including the Robertson Draw Fire which occurred near Red Lodge, were reported in various news outlets, it is reasonable to conclude that the statement is true. Confidence Level: 90%
Demo3		Textual Claim: A photograph depicts a glamorous 'young Harriet Tubman. Prediction: False Explanation: The photograph on the left, which is sometimes claimed to depict a young Harriet Tubman, is actually a misidentified image. The woman in the photograph was actually not Harriet Tubman. Tubman was born in the early 1820s, and there are only a few authenticated photographs of her, with the most famous images showing her as an older woman (as seen on the right). Debunking the myth has involved comparisons of facial features with known photographs of Harriet Tubman as well as historical research into the clothing and photography techniques of the era. Confidence Level: 95%
		Textual Claim: Photographs show thieving birds stealing quarters from a car wash. Prediction: True Explanation: The photograph you provided shows a bird sticking its beak into a coin dispenser, which is a behavior consistent with the claim that birds could be stealing quarters. Birds are known for their intelligence and adaptability, and there are documented cases of birds such as crows using tools and solving problems. It is plausible that a bird learned the association between the coin dispenser and food (since some people might drop food coins there), or they might be attracted to the shiny coins for other reasons such as collecting materials or simply curiosity. Without further context or multiple sources, it's difficult to confirm the intentions of the bird or the frequency of such thievery, but the photo suggests the possibility. Confidence Level: 70%
Demo4		Textual Claim: Photograph shows Hillary Clinton shaking hands with Osama bin Laden. Prediction: False Explanation: The claim that the photograph shows Hillary Clinton shaking hands with Osama bin Laden is a known piece of misinformation. Such a meeting between a high-profile American political figure and the then-leader of Al-Qaeda would have been highly unlikely and well-documented had it happened. Furthermore, the nature of this image, likely appearing on the internet, suggests that it could have been digitally manipulated or a complete fabrication. Without a credible source or confirmation from trustworthy news sources, such a claim should be considered false. Confidence Level: 95%
		Textual Claim: A photograph shows the dramatic differences in two boys who were exposed to the same smallpox source. Prediction: True Explanation: The photograph shown is a classic depiction often used to demonstrate the efficacy of vaccines. The boy on the top appears to have only a few lesions and is in relatively good health, whereas the boy on the bottom is covered with extensive smallpox lesions. This stark contrast is historically used to illustrate the effects of smallpox vaccination. The vaccinated individual (presumably the boy on the top) displays milder symptoms after exposure to the virus, while the unvaccinated individual (presumably the boy on the bottom) shows the full severity of the disease. This image is commonly used in educational materials to highlight the success of vaccination in preventing severe disease. Confidence Level: 90%

Figure 10: Four demonstration pairs used in the ICL experiments, ICL-1 uses only the first example from each demonstration, while ICL-2 uses both examples. They are collected from GPT-4V and factually correct.

		Fauxtography						COSMOS						MOCHEG						Post-4V					
	idx	T	F	U	O	T&F	All	T	F	U	O	T&F	All	T	F	U	O	T&F	All	T	F	U	O	T&F	All
GPT-4V2	0	158	195	29	18	81.9	82.1	179	204	83	14	86.2	80.0	216	223	37	28	87.2	83.5	54	98	26	8	79.6	73.7
		158	181	32	29	84.1	81.3	163	219	75	23	85.1	80.8	205	234	34	31	82.1	81.2	55	95	22	14	77.8	75.3
	5	143	202	31	24	81.2	76.3	175	216	72	17	88.0	81.5	179	250	44	31	78.6	75.6	51	106	18	11	76.4	71.5
LLaVA	0	286	98	3	13	54.4	52.3	331	134	13	2	69.7	67.7	363	112	28	1	61.5	58.1	116	65	2	3	56.3	55.9
	1	197	163	7	33	48.6	48.5	169	236	15	60	69.6	66.3	247	211	9	37	57.6	57.9	68	105	5	8	56.6	57.5
	2	170	223	4	3	49.1	49.5	197	274	7	2	69.0	68.5	318	178	7	1	58.1	57.9	41	141	4	0	56.0	57.0
	3	243	145	8	4	52.6	53.0	243	222	13	2	71.8	70.2	295	197	11	1	57.3	56.3	82	93	10	1	56.5	56.5
	4	211	180	6	3	62.1	62.0	232	244	3	1	77.3	77.1	210	286	4	4	57.3	56.3	55	129	2	0	57.6	57.0
	5	241	154	3	2	63.3	63.0	313	163	4	0	80.5	80.4	260	236	8	0	62.1	62.3	109	75	2	0	52.7	52.7

Table 5: Performance of GPT-4V and LLaVA(13b), with the use of different prompts. Prompt 0 represents the original prompt in Section 3.2; others are suggested by ChatGPT, as shown in Figure 2.

		Fauxtography						COSMOS						MOCHEG						Post-4V					
	Demo Num	T	F	U	O	T&F	All	T	F	U	O	T&F	All	T	F	U	O	T&F	All	T	F	U	O	T&F	All
0	0	286	98	16	0	54.4	52.8	331	134	14	1	69.7	67.7	363	112	29	0	61.5	58.1	116	65	5	0	56.4	55.9
1	1	244	135	9	12	60.9	61.8	295	171	7	7	75.3	74.2	255	229	12	8	62.6	62.7	95	88	3	0	56.3	55.4
	2	182	161	6	51	59.5	60.0	234	220	8	18	75.8	73.5	169	282	10	43	57.4	55.5	42	109	6	29	67.5	63.4
2	1	264	121	4	11	61.6	62.8	337	131	7	5	73.9	73.5	314	180	4	6	64.8	65.1	97	79	10	0	57.4	55.9
	2	176	204	10	10	60.0	60.5	233	230	11	6	80.3	79.6	199	291	8	6	61.2	60.7	44	131	11	0	60.0	60.8
3	1	211	185	1	3	62.6	62.8	279	195	6	0	72.2	71.7	236	266	2	0	61.8	61.5	75	111	0	0	60.2	60.2
	2	191	198	9	2	64.1	63.8	263	201	16	0	75.4	75.4	206	288	9	1	61.1	61.7	54	127	4	1	59.7	59.7
4	1	191	193	9	7	62.2	62.0	260	203	16	1	78.0	77.9	210	283	11	0	60.6	60.9	50	126	8	2	60.8	60.8
	2	196	188	7	9	62.8	62.8	259	207	12	2	77.5	77.5	207	284	11	2	61.1	61.3	52	122	10	2	61.5	61.8

Table 6: Detailed performance of LLaVA(13b) when demonstrations are presented. *Demo*: demonstration index, *Num*: number of examples in the demonstration. The first row, with no demonstration, establishes the baseline. The table presents outcomes from four distinct demonstration pairs, each comprising a true claim followed by a false one.