Can LLMs Improve Multimodal Fact-Checking by Asking Relevant Questions?

Anonymous ACL submission

Abstract

Traditional fact-checking relies on humans to formulate relevant and targeted fact-checking *questions* (FCQs), search for evidence, and verify the factuality of claims. While Large Language Models (LLMs) have been commonly used to automate evidence retrieval and factuality verification at scale, their effectiveness for fact-checking is hindered by the absence of FCQ formulation. To bridge this gap, we seek to answer two research questions: (1) Can LLMs generate relevant FCQs? (2) Can LLM-generated FCQs improve multimodal fact-checking? We therefore introduce a framework LRQ-FACT for using LLMs to generate relevant FCQs to facilitate evidence retrieval and enhance fact-checking by probing information across multiple modalities. Through extensive experiments, we verify if LRQ-FACT can generate relevant FCQs of different types and if LRQ-FACT can consistently outperform baseline methods in multimodal fact-checking. Further analysis illustrates how each component in LRQ-FACT works toward improving the fact-checking performance.

1 Introduction

Fact-checking is an important yet challenging task in combating online misinformation. Modern misinformation often spreads across multiple modalities, containing both textual and visual falsehoods, which significantly complicates accurate and efficient fact-checking (Akhtar et al., 2023). In journalism, fact-checking is traditionally a three-step process (Graves and Amazeen, 2019), where human fact-checkers (1) formulate relevant and targeted fact-checking *questions* (FCQs), (2) search for supporting evidence, and (3) verify the factuality of claims or statements. Fact-checkers leverage domain knowledge to pose precise and contextually relevant FCQs, ensuring claims are evaluated from multiple perspectives (PolitiFact.com, 2011; Vlachos and Riedel, 2014; Vo and Lee, 2019).

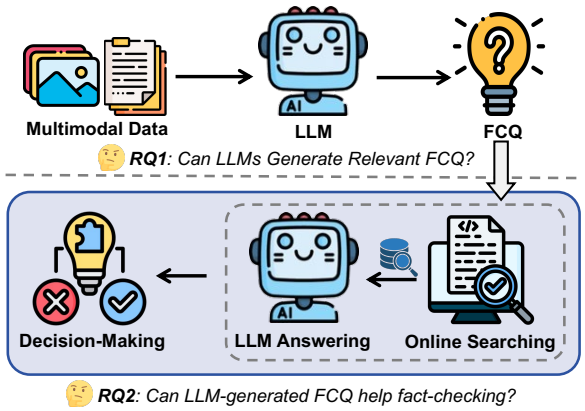

Figure 1: The two research questions we aim to address in this work.

However, given the rapid proliferation of online misinformation (Chen and Shu, 2023; Jiang et al., 2024b), manual fact-checking is insufficient to keep pace with the scale of the problem (Shaeri and Katanforoush, 2023).

To improve efficiency, researchers have developed automated fact-checking (AFC) systems capable of identifying misinformation (Hassan et al., 2017; Miranda et al., 2019; Dierickx et al., 2023). Recently, Large Language Models (LLMs) have been explored for zero-shot fact-checking (Geng et al., 2024). However, research pointed out that directly prompting LLMs for fact-checking remains less effective in many cases (Yao et al., 2023). One key issue is the absence of relevant FCQs, which are essential to guide LLMs in retrieving accurate supporting evidence and conducting reliable veracity evaluations (Chen et al., 2022; Pan et al., 2023; Setty, 2024). In this work, we investigate the potential of LLM-generated FCQs by answering two research questions (Figure 1):

- Are LLMs capable of generating relevant FCQs?
- Can the generated FCQs improve AFC systems?

Inspired by the human fact-checking process, we introduce **L**LM-generated **R**elevant **Q**uestions for multimodal **FACT**-checking (LRQ-FACT), an LLM-based framework designed to automatically

generate relevant and targeted FCQs to guide the AFC system to fact-check multimodal misinformation. LRQ-FACT first generates two types of FCQs: (1) visual FCQs, which assess whether an image accurately represents critical details such as people, objects, or events mentioned in the text, and (2) textual FCQs, which question whether the textual claims or statements are supported by evidence. Human annotators and LLM judges evaluate the quality of LLM-generated FCQs with pre-defined rules, and show that most of the textual and visual FCQs generated by LRQ-FACT are contextually relevant to the fact-checking task.

Next, we seek to answer whether the generated questions can improve multimodal fact-checking. With the guidance of relevant FCQs, LRQ-FACT incorporates the internal training knowledge of LLM and external online searching to strengthen its evidence retrieval and verification capabilities. The up-to-date online information is particularly valuable when fact-checking claims related to emerging or rapidly evolving events, where LLMs often lack sufficient ground truth knowledge. Extensive experiments are conducted on three datasets. Our results show that LRQ-FACT can outperform baseline methods significantly. Furthermore, our ablation study probes into the model's modular components to evaluate their effectiveness in fact-checking performance. Last but not least, we demonstrate that LRQ-FACT is highly adaptive, generalizing across different LLM backbones. In summary, the key contributions of this work are as follows:

- We analyze the use of LLMs to generate relevant and targeted FCQs.
- We explore and experiment the effectiveness of LLM-generated FCQs in multimodal fact-checking on three benchmark datasets.
- Further analysis illustrates how each component in LRQ-FACT contributes to performance.

2 Related Work

2.1 Multimodal Misinformation

Misinformation spans multiple domains, consisting of various modalities such as text and images, making detection increasingly complex (Li et al., 2020; Jiang et al., 2024a; Tufchi et al., 2023). While early misinformation detection mainly focused on textural content (Thorne et al., 2018; Shu et al., 2020), recent datasets have incorporated multimodal misinformation, such as Fakeddit (Nakamura et al., 2019), DGM^4 (Shao et al., 2023), and MMFakeBench (Liu et al., 2024b). Multimodal misinformation detection methods have been developed to learn joint representations of different modalities (Abdali et al., 2022; Zhou et al., 2020). Some studies proposed explainable detection frameworks to enhance the interpretability (Liu et al., 2023; Fung et al., 2021). However, these models may fall short when facing newly emerged or rapidly evolving topics. Our method addresses this issue by utilizing up-to-date Google Search as an external knowledge source.

2.2 Fact-Checking

Fact-checking is essential for combating misinformation, traditionally relying on human fact-checkers to verify claims by generating FCQs and cross-referencing credible sources (Graves, 2018; Graves and Amazeen, 2019). However, manual fact-checking is resource-intensive and struggles to scale with increasing online misinformation (Vlachos and Riedel, 2014). AFC systems address this challenge, with early approaches focusing on textual claims using Machine Learning and Natural Language Processing techniques (Guo et al., 2022; Nakov et al., 2021; Karadzhov et al., 2017). Recent studies have expanded AFC's capabilities by integrating large-scale evidence retrieval (Nie et al., 2019; Akhtar et al., 2023; Geng et al., 2024). However, the lack of FCQs limited the performance of AFC. We address this limitation by harnessing LLMs to generate relevant and targeted FCQs.

2.3 Language Models for Fact-Checking

LLMs and Vision Language Models (VLMs) have had significant impacts on AFC. LLMs from the GPT family (Brown, 2020; Achiam et al., 2023; Hurst et al., 2024) and LLaMA series (Touvron et al., 2023; Vavekanand and Sam, 2024) excel in language understanding and context-aware question generation, enhancing AFC performance (Achiam et al., 2023; Vavekanand and Sam, 2024; Beigi et al., 2024). VLMs such as CLIP (Radford et al., 2021), ViLBERT (Lu et al., 2019), and Paligemma (Beyer et al., 2024), enable cross-modal analysis, aligning visual and textual features to detect false statements. Recent studies highlight the potential of combining LLMs and VLMs to generate targeted FCQs for better fact-checking (Singh et al., 2021; Chen et al., 2022; Pan et al., 2023). Our framework utilizes LLMs and VLMs to generate relevant FCQs, improving multi-modal fact-checking.

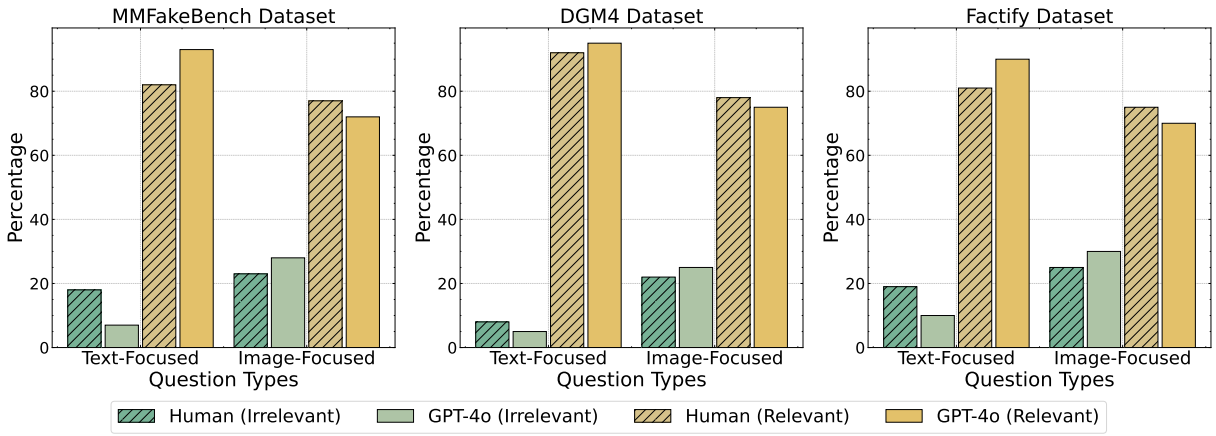

Figure 2: Human and GPT-4o Question Quality Evaluations Across Datasets (50 Samples per Dataset-Modality).

3 Task Definition

We define the task of multimodal fact-checking as a multiclass classification problem. Given a news article $text_i$ and accompanying image img_i as input, we aim to classify the news into one of the following categories:

- **Real** ($y = 0$)**:** The news is factually accurate and consistent.
- **Textual Veracity Distortion (TVD,** $y = 1$)**:** False or misleading claims in the text.
- **Visual Veracity Distortion (VVD,** $y = 2$)**:** Manipulated or misleading images.
- **Cross-Modal Mismatch (CMM,** $y = 3$)**:** Inconsistencies between text and image.

This approach allows for a finer classification of misinformation, improving targeted fact-checking.

4 RQ1: Can LLMs Generate Relevant FCQs?

In this section, we discuss the FCQ generation phase of LRQ-FACT. Specifically, LRQ-FACT produces visual (Sec. 4.1) and textual FCQs (Sec. 4.2). We then evaluate the quality of the generated FCQs using a combination of LLM-based and human assessments (Sec. 4.3).

4.1 Visual FCQs Generation

In the first stage, LRQ-FACT formulates targeted visual FCQs based on the news article to verify whether the visual content aligns with claims in the news article. Importantly, the LLM does not have direct access to the image itself; rather, it generates questions based solely on the article's content, anticipating what aspects might be depicted in an accompanying image. To guide this process, we employ a structured prompt (see Figure 12 in Appendix A.2) that instruct the LLM to focus on objects, settings, interactions, and potential manipulations, to enable a comprehensive verification of accuracy, consistency, and authenticity.

For instance, given a news describing a sporting event, the LLM may generate questions such as:

- *What sport is being played in the image?*
- *Is the pitcher actively throwing the ball?*
- *Are there visible spectators?*
- *Is the image AI-generated or a real photograph?*

4.2 Textual FCQs Generation

The second stage generates fact-checking questions that examine the factual accuracy of textual claims, including dates, names, and events, mimicking human verification methods. Leveraging in-context learning, the LLM formulates precise queries to validate these claims, guided by a structured prompt that ensures relevance and depth in question generation (see Figure 14 in Appendix A.2). For example, if an article states that an umpire performed a "jump action" during a pitch, the LLM might ask:

- *Is it common for an umpire to jump during a pitch in baseball?*
- *Does "launching the ball" align with standard baseball terminology?*
- *Are there baseball rules requiring an umpire to jump while a pitch is thrown?*

4.3 FCQ Quality Evaluation

To ensure a cost-efficient evaluation process, we use human annotators and LLM-as-a-judge to assess the relevance of the generated FCQs. We first establish structured evaluation criteria by incorporating best practices from leading fact-checking organizations (Snopes, 2024; PolitiFact, 2024; FactCheck, 2024). More details can be found in Ta-

ble 4 in Appendix A.4. Specifically, we employ ten predefined criteria to guide the LLM judge in FCQ quality evaluation. These criteria cover aspects such as *critical thinking, analytical depth, precision, factual accuracy, logical consistency, and source credibility* (see Figure 17 in Appendix A.2).

To validate the reliability of the LLM judge, we compare the results of LLM and human annotators (Figure 2) using Fleiss' Kappa correlation (Fleiss, 1971). In particular, we randomly sample 50 instances from each dataset, with each instance containing five visual and five textual FCQs. Two human annotators evaluate these FCQs to decide whether they satisfy all predefined criteria.

Questions \ Dataset	MMFakeBench	DGM4	Factify
Textual	0.78	0.83	0.80
Visual	0.79	0.82	0.81

Table 1: Fleiss' Kappa score between human and LLM evaluations of question relevancy.

Our findings show that LLM-generated FCQs are highly relevant (Table 1), with substantial agreement between human annotators and LLM judge. The high Fleiss' Kappa scores show the potential to extend the LLM-based quality evaluation to all FCQs. As shown in Figure 3, nearly 73% of visual and 93.6% of textual FCQs are relevant. These results highlight the effectiveness of LLMs in generating high-quality FCQs.

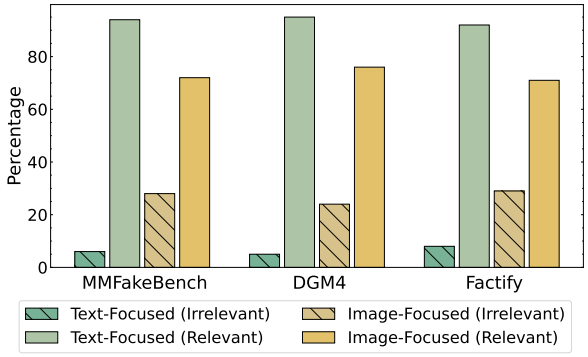

Figure 3: GPT-4o Evaluation of Question Relevance Across Datasets (1000 Samples per Dataset-Modality).

5 RQ2: Can LLM-Generated FCQs improve Multimodal Fact-Checking?

In this section, we first present the remaining components of LRQ-FACT, which includes answering the generated FCQs (Sec. 5.1, 5.2, and 5.3) and a Rule-Based Decision-Maker (Sec. 5.4). We then investigate RQ2 with extensive experiments and analysis on three datasets. We also provide a case study to showcase the effectiveness of FCQs (Sec. 5.8).

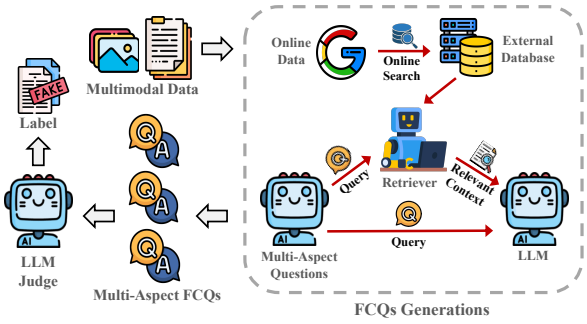

Figure 4: The proposed framework, LRQ-FACT, draws on insights from human fact-checking process.

5.1 Image Description Generation

Since our framework performs the fact-checking in the textual space, we first generate a detailed textual description of the image. The aim is to identify the scene/content/action in the image so that the rule-based stage of the pipeline can use this information to check for consistency between information depicted in the news text vs. shown in the news image. To achieve this, we prompt (see Figure 11 in Appendix A.2) the VLM to generate a summary of the image, instructing it to ensure it to capture key elements of the scene depicted. Formally, given the image img_i, we use the VLM to generate its corresponding description des_i:

$$des_i = \text{VLM}(img_i). \quad (1)$$

5.2 Answering Visual FCQs via VLM

To answer the Visual FCQs, a VLM extracts relevant visual details, allowing for a clear evaluation of how well the image matches the text:

$$ques_i^{v_1}, ...ques_i^{v_m} = \text{LLM}(text_i), \quad (2)$$

$$ans_i^{v_j} = \text{VLM}(img_i, ques_i^{v_j}),\ 0 < j < m. \quad (3)$$

5.3 Answering Textual FCQs via RAG

While LLMs are powerful in language generation, they are prone to hallucinations, leading to inaccurate or unverified information. To enhance factual reliability, we employ Retrieval-Augmented Generation (RAG), which grounds responses in external, verifiable sources. This is especially important for newly emerging topics in news articles since the LLM may not have knowledge of such topics due to its knowledge cutoff. We implement an automated online search by using the Google Web Search

API (Google) to gather relevant news articles for each claim. We identify the top 10 sources, extract their textual content, and compile a factual document containing relevant information. Next, we use LangChain (Chase, 2022) to retrieve the most relevant passages from this document for fact-checking questions, ensuring that only contextually accurate information is used for verification. Finally, an LLM generates answers based on the retrieved content, reducing hallucinations and improving factual accuracy. To guide the LLM effectively, we employ a carefully designed prompt (see Figure 16 in Appendix A.2) that instructs the model to rely solely on the provided factual evidence. This process is formulated as follows:

$$Doc_i = \text{SearchOnline}(text_i) \tag{4}$$

$$\begin{aligned} RelContent_i &= \text{Retriever}(Doc_i, ques_i^{t_j}) \\ ans_i^{t_j} &= \text{LLM}(RelContent_i, ques_i^{t_j}) \\ & 0 < j < n. \end{aligned} \tag{5}$$

5.4 Rule-Based Decision-Maker

To enhance the judge LLM's ability to effectively utilize multimodal evidence and make accurate predictions, we introduce a rule-based decision-maker module. This module guides the judge LLM to follow expert-like reasoning steps:

- **General Instructions:** We provide guidelines to help judge LLMs establish connections between QA analyses and specific labels. For example, if evidence from FCQs contradicts a claim, the decision-maker assigns it the label "TVD".
- **Additional Guidelines:** We outline supplementary instructions, such as analyzing facial expressions, identifying unrealistic elements, and verifying cross-modal consistency, enabling judge LLMs to conduct a human-like examination.
- **Output Format:** We specify the required output format, including the judgment, a fine-grained misinformation label, and a detailed explanation.

For each piece of analyzed content, the decision-maker provides: (1) a final judgment $j_i \in$ {Real, Fake}, (2) a label $l_i \in$ {Textual Veracity Distortion, Visual Veracity Distortion, Mismatch} identifying the specific issue, and (3) a detailed explanation of the decision-making process e_i. This rule-based approach ensures that the framework provides clear, evidence-based conclusions, carefully weighing the alignment between textual and visual information to detect misinformation. This process can be formulated as:

$$qa_i^v = \oplus_{j=1}^{m}[ques_i^{v_j}, ans_i^{v_j}], \tag{6}$$

$$qa_i^t = \oplus_{j=1}^{n}[ques_i^{t_j}, ans_i^{t_j}], \tag{7}$$

$$j_i, l_i, e_i = \text{LLM}(text_i, img_i, des_i, qa_i^v, qa_i^t), \tag{8}$$

where $\oplus$ denotes the concatenation operation, and m and n represent the number of visual and textual FCQs, respectively.

5.5 Datasets

MMFakeBench (Liu et al., 2024b) contains 11,000 image-text pairs. It goes beyond the assumption of single-source forgery and presents samples with *"Real"*, *"Textual Veracity Distortion"*, *"Visual Veracity Distortion"*, and *"Cross-modal Consistency Distortion"*, with both human- and machine-generated images.

DGM^4 (Shao et al., 2023) is a large-scale multimodal dataset comprising 230,000 image-text paired samples. Image manipulation in the dataset involves *"face swapping and facial emotion editing"*, while text manipulation includes *"sentence replacement and textual sentiment editing"*. The DGM^4 dataset is constructed based on the VisualNews dataset (Liu et al., 2020), which collects data from multiple news agencies.

Factify (Mishra et al., 2022) is a multimodal fact-checking benchmark comprising 50,000 data points, each consisting of a textual claim, an associated image, and corresponding reference documents. It is categorized into three main classes: *"Support"*, *"NotEnoughInfo"*, and *"Refute"*, with finer-grained labels for detailed evaluation.

More details about the datasets are provided in Appendix A.5.

5.6 Experiment Details and Settings

For MMFakeBench and DGM^4, we sample 1,000 validation instances, ensuring a balanced distribution: 300 *Real*, 300 *TVD*, 100 *VVD*, and 300 *CMM*. DGM^4 labels are mapped as follows: *real* ("orig"), *VVD* ("face swap," "face attribute"), *TVD* ("text attribute"), and *mismatch* ("text swap") to ensure consistency across datasets. For Factify, we sample 750 validation instances: 300 Support (*Support Multimodal*, *Support Text*), 300 NotEnoughInfo (*Insufficient Text*, *Insufficient Multimodal*), and 150 Refute, maintaining a balanced evaluation.

Backbone	Approach	MMFakeBench		DGM4		Factify	
		F1↑	ACC↑	F1↑	ACC↑	F1↑	ACC↑
VILA (Lin et al., 2024)	SP	11.5	30.0	19.4	19.8	25.6	27.7
InstructBLIP (Dai et al., 2023)	SP	13.7	28.8	19.2	19.8	24.3	26.9
BLIP-2 (Li et al., 2023)	SP	16.7	32.8	18.1	19.1	26.6	28.6
LLaVA-1.6 (Liu et al., 2024a)	SP	25.7	40.4	32.5	39.4	51.3	56.2
GPT-4V-1.7T (OpenAI, 2023)	SP	51.0	54.0	42.3	51.5	64.2	68.2
GPT-4o (OpenAI, 2023)	SP	49.2	60.9	39.9	55.9	72.5	71.2
LRQ-FACT (w/o RAG)	FCQs	66.5	65.5	45.8	58.0	-	-
LRQ-FACT (w/ RAG)	FCQs	**71.6**	**70.8**	**49.2**	**62.3**	**75.2**	**73.1**

Table 2: The last two rows represent the results of LRQ-FACT. Standard prompt (SP) refers to a generic prompt without fact-checking questions. **Bold**: best result; Underline: second best result.

LRQ-FACT integrates both LLMs and VLMs to handle the tasks of question generation and answering. For generating visual and textual questions, we use the GPT-4o (Achiam et al., 2023), along with Paligemma (Beyer et al., 2024) and LLaMA 3.1 (Vavekanand and Sam, 2024) as different backbones for VLMs and LLMs respectively. For each modality, we generate five questions by employing in-context learning, where 10 example questions are provided for each modality to guide the model in generating high-quality and relevant questions. To answer fact-checking questions, we employ two settings: (1) using only LLM knowledge and (2) a RAG approach, where an external knowledge retrieval module fetches relevant supporting documents before generating responses. In the decision-making phase, GPT-4o is also used to leverage the FCQs for fact verification. Known for its strong reasoning and rule-following capabilities, GPT-4o assesses the veracity of the content and offers clear rationales for its conclusions. All experiments are conducted on NVIDIA 40GB V100 GPUs.

5.6.1 Evaluation Metrics

To assess the performance of the various baselines, we employ a multi-class classification approach. Following the practices established in prior works (Zhang and Gao, 2023; Qian et al., 2021; Chen et al., 2023), we use the widely adopted macro-F1 score as our primary evaluation metric. The macro-F1 score provides a balanced measure of precision and recall through their harmonic mean, ensuring fair evaluation across all classes. In addition to the macro-F1 score, we also report macro-accuracy as complementary metrics, offering a more comprehensive understanding of model performance.

5.6.2 Comparison Models

We select a diverse range of VLMs as baseline models for comparison. These include: InstructBLIP (Dai et al., 2023), VILA (Lin et al., 2024), BLIP-2 (Li et al., 2023), LLaVA-1.6 (Liu et al., 2024a) and the closed-source GPT-4V-1.7T, GPT-4o (OpenAI, 2023) models.

5.7 Experiment Analysis

To assess whether LLM-generated FCQs enhance multimodal fact-checking, we conduct a three-stage analysis: (1) a comparison between our LRQ-FACT and other baseline methods, (2) an ablation study to measure the impact of generated questions on fact-checking performance and (3) an in-depth examination of how different types of questions contribute to improving verification accuracy across modalities.

5.7.1 Comparison with Other Methods

Table 2 compares our LRQ-FACT model with baseline methods. Our proposed approach consistently outperforms all baselines across various datasets, regardless of whether external knowledge is introduced (w/ RAG) or not (w/o RAG). This highlights the effectiveness of incorporating FCQs in enhancing fact-checking performance.

5.7.2 Effect of LLM-Generated FCQs on Fact-Checking Performance

Table 3 reports F1 scores and accuracy (ACC) for various ablation configurations, comparing the baseline GPT-4o model with visual and textual FCQs variations.

Overall Improvements: Incorporating FCQs enhances fact-checking performance across all

Method	MMFakeBench		DGM4		Factify	
	F1	ACC	F1	ACC	F1	ACC
LRQ-FACT (VLM: Paligemma & LLM: LLaMA 3.1)	51.2	56.4	41.2	53.6	66.2	65.3
w/ Textual FCQs	$62.5_{+22.1\%}$	$59.1_{+4.8\%}$	$43.3_{+5.1\%}$	$54.1_{+0.9\%}$	$64.8_{-2.1\%}$	$60.9_{-6.7\%}$
w/ Visual FCQs (w/o RAG)	$59.4_{+16.0\%}$	$57.3_{+1.6\%}$	$46.1_{+11.8\%}$	$53.9_{+0.5\%}$	-	-
w/ Visual & Textual FCQs (w/o RAG)	$62.8_{+22.6\%}$	$61_{+8.1\%}$	$47.7_{+15.7\%}$	$54.8_{+2.2\%}$	-	-
w/ Textual FCQs (w/ RAG)	$61.7_{+20.5\%}$	$63.2_{+12.1\%}$	$48.3_{+17.2\%}$	$61.7_{+15.1\%}$	$69.5_{+5.0\%}$	$67.2_{+2.9\%}$
w/ Visual & Textual FCQs (w/ RAG)	$64.2_{+25.3\%}$	$64.8_{+14.8\%}$	$48.6_{+17.9\%}$	$62.1_{+15.8\%}$	$73.2_{+10.5\%}$	$71.5_{+9.4\%}$
LRQ-FACT (VLM & LLM: GPT-4o)	49.2	60.9	39.9	55.9	72.5	71.2
w/ Textual FCQs	$64.3_{+30.7\%}$	$62.1_{+2.0\%}$	$40.4_{+1.3\%}$	$54.3_{-2.9\%}$	$71.1_{-1.9\%}$	$68.7_{-3.5\%}$
w/ Visual FCQs (w/o RAG)	$59.6_{+21.1\%}$	$61.7_{+1.3\%}$	$47.2_{+18.3\%}$	$59.7_{+6.8\%}$	-	-
w/ Visual & Textual FCQs (w/o RAG)	$\underline{66.5}_{+35.2\%}$	$\underline{65.5}_{+7.5\%}$	$45.8_{+14.8\%}$	$58.0_{+3.8\%}$	-	-
w/ Textual FCQs (w/ RAG)	$61.8_{+25.2\%}$	$64.7_{+6.2\%}$	$\mathbf{49.4}_{+23.8\%}$	$\mathbf{62.5}_{+11.8\%}$	$\underline{74.0}_{+2.1\%}$	$\underline{72.5}_{+1.8\%}$
w/ Visual & Textual FCQs (w/ RAG)	$\mathbf{71.6}_{+45.5\%}$	$\mathbf{70.8}_{+16.3\%}$	$\underline{49.2}_{+23.3\%}$	$\underline{62.3}_{+11.5\%}$	$\mathbf{75.2}_{+3.7\%}$	$\mathbf{73.1}_{+2.7\%}$

Table 3: Ablation study result. The subscript values indicate the percentage improvement over the GPT-4o baseline.

datasets. For MMFakeBench, the F1 score improves from 49.2 to 71.6 (+45.5%), and accuracy rises from 60.9 to 70.8 (+16.3%). Similarly, for DGM^4, the highest F1 score increases from 39.9 to 49.2 (+23.3%), while accuracy improves from 55.9 to 62.3 (+11.5%). Factify also benefits from FCQs integration, with F1 increasing from 72.5 to 75.2 (+3.7%) and accuracy from 71.2 to 73.1 (+2.7%).

Impact of Visual FCQs: Visual FCQs alone yield noticeable improvements in both F1 and accuracy across all datasets. For MMFakeBench, F1 improves by 30.7% (from 49.2 to 64.3), while accuracy rises slightly (+2.0%). A similar trend is observed for DGM^4, where F1 improves by 1.3% and accuracy by 2.9%. This improvement can be attributed to the fact that visual FCQs encourage the model to focus more on visual details, prompting a deeper analysis of the image content.

Impact of Textual FCQs without RAG: The addition of textual FCQs without explicit external evidence shows significant gains, especially in DGM^4, where the F1 score rises from 39.9 to 47.2 (+18.3%), and accuracy improves by 6.8%. However, this setting is not evaluated for Factify because the task in this dataset is verification rather than detection, and Factify already provides factual reference documents for verification. This improvement can be attributed to the fact that textual FCQs systematically capture all claims within the news and probe their validity. By generating targeted questions, the model breaks down the text into specific factual assertions, allowing for a more structured verification process. This focused approach helps detect inconsistencies, misinterpretations, or misleading statements within the text, ultimately leading to more accurate fact-checking.

Impact of Textual FCQs with RAG: When textual FCQs incorporate external evidence, fact-checking performance further improves. For MMFakeBench, F1 increases to 61.8 (+25.2%), and accuracy improves by 6.2%. In DGM^4, this approach achieves the highest F1 (49.4, +23.8%) and accuracy (62.5, +11.8%), highlighting the importance of retrieved evidence. Factify also benefits, with F1 rising to 74.0 (+2.1%) and accuracy to 72.5 (+1.8%).

These results confirm that integrating LLM-generated FCQs, particularly when paired with external evidence, significantly enhances multimodal fact-checking performance.

5.7.3 Effect of Fact-Checking Questions on Fact-Checking Performance

To further investigate how different types FCQs impact verification performance, we analyze recall scores across different manipulation categories within the MMFakeBench dataset. Figure 5 presents recall values for various settings, including visual FCQs (VFCQs), textual FCQs (TFCQs), and combinations of both with and without RAG.

Impact on Real and Mismatch Cases: While FCQs improve fact-checking performance overall, recall scores for the *Real* and *Mismatch* categories decrease. This phenomenon occurs because, in the absence of structured questioning, the model does not apply strict fact-checking criteria and tends to classify most news samples as either *Real* or *Mismatch*. However, with FCQs, the model adopts a more expert-like approach, becoming more cautious before confirming a claim as real. This in-

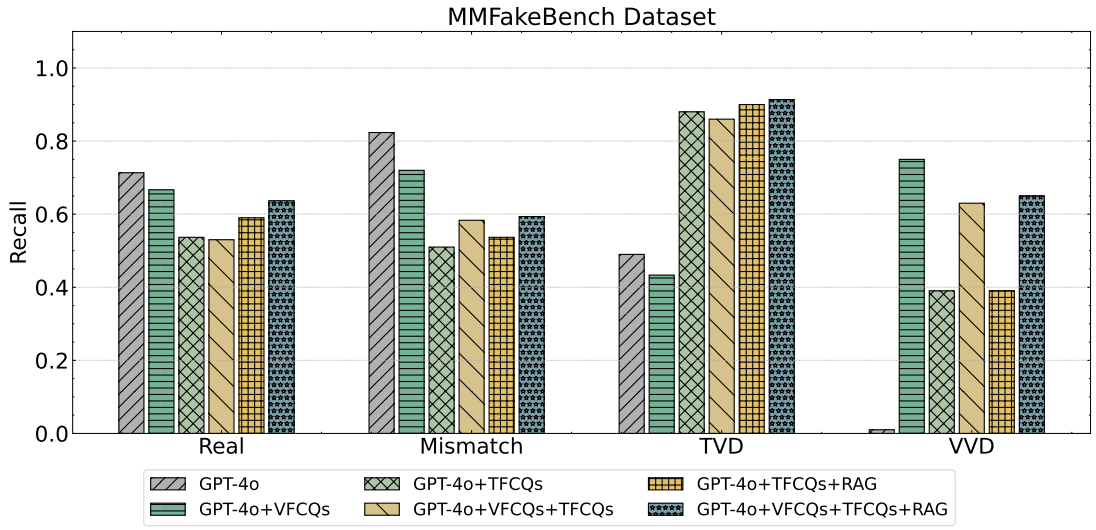

Figure 5: Detailed ablation study result. TFCQs and VFCQs represent Textual FCQs, Visual FCQs respectively.

creased scrutiny aligns with human verification processes, reducing false positives but leading to a lower recall in these categories.

Impact on Textual Veracity Distortion (TVD): FCQs significantly enhance recall for cases involving textual manipulation. Without structured questioning, the model relies primarily on its internal knowledge, which may not always be up-to-date or factually accurate. However, incorporating textual FCQs (TFCQs) allows the model to engage in a more structured verification process by utilizing factual information instead of relying solely on pre-trained LLM knowledge. Additionally, integrating external evidence (TFCQs+RAG) enhances the model's ability to detect inconsistencies in manipulated text, improving misinformation identification. This highlights the value of supplementing LLM-generated responses with retrieved factual data for more reliable fact-checking.

Impact on Visual Veracity Distortion (VVD): Visually manipulated claims pose a challenge for the baseline model, which often lacks the ability to rigorously assess image alterations. However, incorporating visual FCQs (VFCQs) substantially improves recall by prompting the model to analyze visual content more critically. The highest recall is achieved when both VFCQs and TFCQs are used together with external evidence, reinforcing the importance of cross-modal verification.

Across all manipulation types, the most effective setup involves combining VFCQs and TFCQs with external evidence (VFCQs+TFCQs+RAG). While this structured questioning approach makes the model more cautious in verifying real claims, it significantly enhances misinformation detection, ensuring a more reliable fact-checking process. These findings highlight how FCQs improve fact-checking by encouraging a more rigorous verification process. By systematically questioning claims across both textual and visual modalities and incorporating factual retrieval mechanisms, FCQs help the model adopt a more expert-like approach, leading to more precise and reliable fact verification.

5.8 Case Study

To demonstrate the effectiveness of FCQs in detecting and classifying multimodal misinformation, we provide examples in the Appendix (see Figures 7, 8, 9, and 10) showcasing their impact.

6 Conclusion

In this work, we investigate whether LLMs can generate relevant FCQs and whether these LLM-generated FCQs can improve multimodal fact-checking. Through the proposed framework, LRQ-FACT, we demonstrate that LLMs are indeed capable of generating highly relevant and targeted FCQs, effectively addressing a key limitation in AFC systems. Furthermore, our experiments show that incorporating relevant FCQs into the fact-checking process significantly enhances evidence retrieval and improves the overall factuality verification performance. LRQ-FACT outperforms baseline methods, showcasing the effectiveness of FCQ generation in strengthening multimodal fact-checking. These results highlight the potential of LLM-based FCQ formulation as a promising direction in future AFC research, facilitating more reliable and scalable methods.

Limitations

While our study demonstrates that LLM-generated FCQs enhance multimodal fact-checking, several limitations remain. One major limitation is the lack of expert-level validation. Although we evaluate FCQ quality using benchmark datasets and human annotations, we do not assess whether the generated questions exhibit reasoning comparable to domain experts. Incorporating expert evaluations, particularly in specialized fields such as medicine or law, could provide a more rigorous assessment of FCQ quality and alignment with high-quality fact-checking standards.

Another limitation is the absence of a random question baseline. Our experiments compare LRQ-FACT against strong fact-checking baselines, but we do not explicitly test whether randomly generated questions could serve as a control. Introducing a random baseline would help isolate the actual contribution of meaningful FCQ generation from the broader effect of question-driven retrieval. Additionally, our approach is inherently dependent on the capabilities of the underlying LLM. If the model produces vague, misleading, or hallucinated questions, it could negatively impact fact-checking performance by retrieving irrelevant or incorrect evidence. Further investigation into fine-tuned models or more controlled question-generation strategies could mitigate these risks.

Our evaluation process also introduces certain limitations. The FCQs were assessed by PhD students who received detailed instructions and predefined criteria to evaluate question relevance. While this ensures a structured and consistent evaluation process, the annotator pool is relatively small and may not fully represent diverse perspectives. The background knowledge of annotators could influence their judgments, and a broader demographic, including professional fact-checkers or domain experts, may provide a more comprehensive evaluation of FCQ effectiveness.

The reliance on external search engines for evidence retrieval introduces another source of variability. The effectiveness of our approach depends on search engine algorithms, indexing policies, and the availability of high-quality sources, which may not always be consistent. This issue is particularly relevant for fact-checking emerging claims or topics where authoritative sources are scarce. Furthermore, while LRQ-FACT integrates textual and visual evidence for multimodal fact-checking, its reliance on image captions and textual representations of visual content may introduce errors. Exploring direct visual analysis through vision-language models or image embeddings could improve robustness in cases where textual descriptions are insufficient.

Generalization across fact-checking domains remains another open challenge. Although our approach performs well on benchmark datasets, its effectiveness across diverse domains such as political misinformation, scientific fact-checking, and real-time verification is not fully explored. Future work should investigate domain-specific FCQ generation techniques to adapt the framework for specialized fact-checking tasks. Additionally, the computational overhead of our approach may limit its practical deployment. Multiple inference steps, including FCQ generation, evidence retrieval, and multimodal reasoning, contribute to significant resource consumption, making real-time fact-checking more challenging. Optimizing the framework for efficiency would be necessary for large-scale deployment.

Finally, our evaluation process relies on both human annotators and LLM-based scoring to assess FCQ quality. While high agreement between human and model-based evaluations suggests reliability, potential biases in human annotation criteria and systematic artifacts in LLM-generated scoring may still influence results. Future research should explore alternative evaluation metrics and methodologies to ensure robustness and fairness in assessing FCQ effectiveness.

Ethical Statement

This work explores the use of LLMs to generate fact-checking questions (FCQs) for automated verification. While our approach enhances misinformation detection, it raises ethical considerations. LLM-generated FCQs may reflect biases present in training data, potentially influencing fact-checking outcomes. Although we evaluate FCQ quality with human reviewers and benchmark datasets, further efforts are needed to ensure fairness and neutrality.

Another concern is the risk of over-reliance on automation. While LLMs can support fact-checking at scale, they should not replace human judgment, particularly in high-stakes domains like politics and healthcare. Our framework is designed as an assistive tool, reinforcing rather than replacing expert oversight. Privacy considerations are

also critical, as our approach retrieves publicly available evidence without processing personally identifiable information. We adhere to ethical data usage practices and fair use policies but recognize the need for continuous alignment with evolving privacy standards.

Finally, transparency and accountability in AI-driven fact-checking remain essential. Black-box decision-making can undermine trust, emphasizing the importance of explainability. By acknowledging these challenges, we advocate for responsible AI deployment, promoting fairness, human oversight, and transparency in automated fact-checking systems.

y whether the questions exhibit "expert-like" reasoning. Annotations from domain experts could provide a more rigorous evaluation of FCQ quality, particularly in specialized fields like medical or scientific fact-checking.

References

Sara Abdali, Bhaskar Krishnamachari, et al. 2022. Multi-modal misinformation detection: Approaches, challenges and opportunities. *arXiv preprint arXiv:2203.13883*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal automated fact-checking: A survey. *arXiv preprint arXiv:2305.13507*.

Alimohammad Beigi, Zhen Tan, Nivedh Mudiam, Canyu Chen, Kai Shu, and Huan Liu. 2024. Model attribution in machine-generated disinformation: A domain generalization approach with supervised contrastive learning. *arXiv preprint arXiv:2407.21264*.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Harrison Chase. 2022. LangChain.

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. *arXiv preprint arXiv:2205.06938*.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 627–638.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Laurence Dierickx, Carl-Gustav Lindén, and Andreas Lothe Opdahl. 2023. Automated fact-checking to support professional practices: systematic literature review and meta-analysis. *International Journal of Communication*, 17:21–21.

FactCheck. 2024. Factcheck.org - monitoring factual accuracy in u.s. politics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avirup Sil. 2021. Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698.

Jiahui Geng, Yova Kementchedjhieva, Preslav Nakov, and Iryna Gurevych. 2024. Multimodal large language models to support real-world fact-checking. *arXiv preprint arXiv:2403.03627*.

Google. The google web apis.

D Graves. 2018. Understanding the promise and limits of automated fact-checking. *Reuters Institute for the Study of Journalism*.

Lucas Graves and Michelle Amazeen. 2019. Fact-checking as idea and practice in journalism.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD*

international conference on knowledge discovery and data mining, pages 1803–1812.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Bohan Jiang, Zhen Tan, Ayushi Nirmal, and Huan Liu. 2024a. Disinformation detection: An evolving challenge in the age of llms. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, pages 427–435. SIAM.

Bohan Jiang, Chengshuai Zhao, Zhen Tan, and Huan Liu. 2024b. Catching chameleons: Detecting evolving disinformation generated using large language models. *arXiv preprint arXiv:2406.17992*.

Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. Fully automated fact checking using external sources. *arXiv preprint arXiv:1710.00341*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. *arXiv preprint arXiv:2011.04088*.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pretraining for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2020. Visual news: Benchmark and challenges in news image captioning. *arXiv preprint arXiv:2010.03743*.

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

Hui Liu, Wenya Wang, and Haoliang Li. 2023. Interpretable multimodal misinformation detection with logic reasoning. *arXiv preprint arXiv:2305.05964*.

Xuannan Liu, Zekun Li, Peipei Li, Shuhan Xia, Xing Cui, Linzhi Huang, Huaibo Huang, Weihong Deng, and Zhaofeng He. 2024b. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. *arXiv preprint arXiv:2406.08772*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Sebastião Miranda, David Nogueira, Afonso Mendes, Andreas Vlachos, Andrew Secker, Rebecca Garrett, Jeff Mitchel, and Zita Marinho. 2019. Automated fact checking in the news room. In *The world wide web conference*, pages 3579–3583.

Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya N Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit P Sheth, Asif Ekbal, et al. 2022. Factify: A multi-modal fact verification dataset. In *DE-FACTIFY@ AAAI*.

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

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. *arXiv preprint arXiv:2103.07769*.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6859–6866.

GPT OpenAI. 2023. 4v (ision) system card. *preprint*.

Liangming Pan, Xinyuan Lu, Min-Yen Kan, and Preslav Nakov. 2023. Qacheck: A demonstration system for question-guided multi-hop fact-checking. *arXiv preprint arXiv:2310.07609*.

PolitiFact. 2024. Politifact - fact-checking u.s. politics.

PolitiFact.com. 2011. Principles of politifact and the truth-o-meter. Accessed: 2024-10-04.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5434–5445.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Vinay Setty. 2024. Surprising efficacy of fine-tuned transformers for fact-checking over larger language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2842–2846.

Pouya Shaeri and Ali Katanforoush. 2023. A semi-supervised fake news detection using sentiment encoding and lstm with self-attention. In *2023 13th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 590–595. IEEE.

Rui Shao, Tianxing Wu, and Ziwei Liu. 2023. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Vivek K Singh, Isha Ghosh, and Darshan Sonagara. 2021. Detecting fake news stories via multimodal analysis. *Journal of the Association for Information Science and Technology*, 72:3–17.

Snopes. 2024. Snopes - the definitive fact-checking resource.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Shivani Tufchi, Ashima Yadav, and Tanveer Ahmed. 2023. A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities. *International Journal of Multimedia Information Retrieval*, 12(2):28.

Raja Vavekanand and Kira Sam. 2024. Llama 3.1: An in-depth analysis of the next-generation large language model.

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

Nguyen Vo and Kyumin Lee. 2019. Learning from fact-checkers: analysis and generation of fact-checking language. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–344.

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*, pages 2733–2743.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1011.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049.

A Appendix

A.1 Example Analysis

Real. In this case study (Figure 7), the description is well-constructed and aligns perfectly with the image and the textual context. The depiction of Jake Davis as a young man in casual clothing, standing in a relaxed manner, accurately reflects the narrative of his release from a young offender institution. The visual context provided in the image adds credibility to the news article, confirming the validity of the description. There are no discrepancies between the image and the text, making the description not only good but also a reliable tool to confirm the factual correctness of the news.

The questions presented in this case are well-formed and reliable. They are designed to extract key details from both the image and the text, ensuring comprehensive verification. The visual questions effectively ask about the setting and identity, which helps in confirming whether the person and location in the image match the article's claims. The text-based questions aim to validate the timeline and factual details, ensuring a consistent narrative. These questions are precise and structured to get the best possible answers, making them a solid mechanism for cross-verifying facts.

Textual Veracity Distortion. The description in Figure 8 accurately depicts a lighthouse in a Gothic architectural style, positioned on a rocky shore with surrounding water and seagulls. The image is valid and corresponds with the article's general theme. However, the description's alignment with the actual claim in the text—that the lighthouse is haunted and located in Greece—proves to be incorrect. While the description is visually consistent and good, it does not support the erroneous textual claim, showing how important it is to assess both text and visuals in tandem.

The questions in this case are reliable and appropriately structured to identify discrepancies between the image and the text. The visual questions ask about the architectural style and

contextual clues from the image, while the text-based questions explore the factual accuracy of the claim that this lighthouse is haunted and located in Greece. The questions provide a good framework for fact-checking by encouraging thorough scrutiny of both visual and textual elements. This ensures that any distortions or misrepresentations in the article are effectively highlighted, making the questions a valuable tool for getting to the truth.

Visual Veracity Distortion. In Figure 9, the description of a clock tower in yellow and white is valid and clear, but the image itself shows a structure that is clearly gold and digitally altered. The description is good in terms of clarity and helping readers visualize the article's claim, even though it does not reflect the manipulated nature of the image. This highlights the importance of analyzing the veracity of visuals alongside textual descriptions.

The questions are well-crafted to reveal any visual inconsistencies. The visual questions ask about the color and reality of the clock tower, which are key to identifying that the clock tower has been digitally altered. The text-based questions, which probe the existence of such a clock tower in real life, also help uncover discrepancies. These questions are reliable and precise, aimed at extracting the best possible answers and guiding the evaluation of the article's claims against the evidence provided by the image.

Cross-modal Consistency Distortion. This case (Figure 10) involves a clear mismatch between the text and the image, where the article describes a little girl holding uncooked rolls, while the image shows her holding paper towels. The description is coherent and well-explained, making it a good tool to visualize the scenario presented in the article. However, the inconsistency between the image and the text highlights a cross-modal distortion. Despite this, the description itself remains valid in its own right.

The questions presented are well-designed to highlight the inconsistency between the text and the image. The visual questions ask about the scene and the object the girl is holding, providing clear answers that reveal the mismatch. The text-based questions further confirm this by addressing the article's lack of accurate description. These questions are well-structured and reliable, allowing for an in-depth examination of both the image and the text to expose cross-modal discrepancies. They guide the analysis toward the best possible answers by focusing on the key elements that need verification.

A.2 Instruct Prompt for LRQ-FACT

The LRQ-FACT framework employs a series of structured prompts to guide LLMs and VLMs in multimodal fact-checking. These prompts facilitate the generation of detailed image descriptions, contextually relevant questions, and well-informed answers that probe the veracity of both visual and textual content. In the final step, a rule-based decision-maker evaluates the generated questions and answers to provide a final judgment on the consistency between the text and image, ensuring accurate detection of misinformation.

Image Description Prompt. The first step is to generate a detailed description of the image, capturing all relevant elements that help assess its consistency with the textual content. This description is crucial for identifying potential inconsistencies or manipulations between the image and the accompanying article. The specific prompt used to generate this description is provided in Figure 11.

Visual Questions Prompt. This stage generates relevant visual questions designed to verify the accuracy, authenticity, and relevance of the visual content in relation to the article. These questions help clarify the image content and assess its relation to the text. The specific prompt for generating these questions is illustrated in Figure 12.

Visual Answers Prompt. After generating the visual questions, this prompt helps in generating answers that analyze the visual content directly from the image. These answers are based on the key elements and actions identified in the image, ensuring that the responses are relevant and insightful. The specific prompt for this is shown in Figure 13.

Textual Questions Prompt. To critically assess the factual claims in the text, this prompt generates relevant questions targeting specific elements such as dates, names, locations, and events. The generated questions aim to challenge the accuracy of the claims made in the article. The specific prompt used for textual questions is shown in Figure 14.

Textual Answers Prompt. After generating textual questions, this prompt enables the model to generate answers using its built-in knowledge. The specific prompt is shown in Figure 15. Additionally, we employ a Retrieval-Augmented Generation (RAG) approach to incorporate factual evidence, ensuring more reliable and verifiable re-

sponses. These answers help assess factual accuracy and challenge any unsupported claims in the article. The corresponding RAG-based prompt is illustrated in Figure 16.

Question Quality Assessment Prompt. To evaluate the relevance of generated questions, we use a fact-checking criteria-based prompt that classifies questions as relevant or irrelevant. This assessment considers factors such as alignment with the claim, specificity, and usefulness in verifying factual accuracy. The specific prompt used for this evaluation is illustrated in Figure 17.

Rule-Based Decision-Maker Prompt. After gathering information from the image and text analyses, the rule-based decision-maker evaluates the consistency between modalities and makes a final determination about the article's veracity. This module provides a detailed explanation for the final judgment. The specific prompt for the rule-based decision-making process is shown in Figure 18.

A.3 Annotator Details

To evaluate the quality of LLM-generated FCQs, we recruited two PhD students with backgrounds in NLP and computational linguistics. Annotators were provided with detailed instructions and predefined criteria to assess the relevance of each FCQ to the given claim. The evaluation process aimed to ensure consistency and minimize subjectivity in judgment. While this setup provides structured and knowledgeable assessments, the annotator pool is relatively small and may not fully capture diverse perspectives. Future work could incorporate domain experts or professional fact-checkers to further validate FCQ effectiveness across different fact-checking domains.

A.4 Criteria for Evaluating FCQ Quality

To systematically evaluate the quality of LLM-generated fact-checking questions (FCQs), we developed a structured evaluation framework inspired by best practices from established fact-checking methodologies. Our evaluation process consists of two key components: LLM-based assessment and human evaluation, ensuring a rigorous and reliable analysis of question relevance.

Evaluation Framework. We designed our evaluation framework to assess the effectiveness of both *visual* and *textual* FCQs. The framework follows ten evaluation criteria, derived from widely accepted fact-checking principles, emphasizing accuracy, credibility, and relevance. These criteria help determine whether the generated FCQs effectively probe factual claims and align with real-world verification standards (Figure 17).

One challenge in evaluating FCQs is ensuring question specificity without over-constraining the verification process. A well-formed FCQ should allow multiple valid answers depending on available evidence while still prompting meaningful fact-checking efforts. Additionally, cross-modal consistency is a key factor in multimodal fact-checking—image-based questions must align with textual claims without introducing unintended biases or assumptions.

LLM-Based vs. Human Evaluation. To ensure consistency, we employ GPT-4o as an automated evaluator, scoring FCQs based on predefined criteria such as logical structure, factual precision, and investigative depth. However, LLM-based evaluations may still miss nuanced contextual ambiguities that a human fact-checker would recognize, such as misleading phrasing or assumptions embedded in a question.

To validate the reliability of the LLM-based assessment, we conducted a human agreement study, comparing GPT-4o's evaluation results with expert annotations across datasets in Table 1. The goal was to determine the degree of alignment between human and LLM judgments, rather than integrating both assessments into a single process.

Our findings indicate that while LLMs are effective at systematically evaluating FCQs, human reviewers provide valuable qualitative insights, particularly in identifying question formulation errors that could lead to misinformation rather than prevent it.

This subtle difference can impact both retrieval accuracy and the framing of fact-checking results.

Insights from the Evaluation Process.

- **Text-based FCQs generally receive higher relevance scores than image-based FCQs.** This discrepancy suggests that LLMs have a better grasp of linguistic verification than visual reasoning, which remains an open challenge in multimodal misinformation detection.
- **Human annotators tend to be stricter in rejecting vague or broad FCQs.** LLM-based evaluations show slightly higher acceptance rates for questions that are loosely related to the claim but lack clear fact-checking intent.
- **Context-aware evaluation is critical.** Without access to real-world updates, an FCQ might appear factually valid but be outdated or misleading

in light of new developments. This highlights the importance of external knowledge retrieval in automated fact-checking pipelines.

By comparing human and LLM-based assessments, our study confirms that GPT-4o produces highly relevant FCQs with near-human accuracy. However, human reviewers remain essential in refining question design and identifying subtle logical inconsistencies that automated evaluations may overlook.

A.5 Dataset Descriptions and Details

To assess the effectiveness of LLM-generated fact-checking questions in multimodal misinformation detection, we utilize three benchmark datasets: MMFakeBench, DGM4, and Factify. These datasets encompass a wide range of real and manipulated image-text pairs, enabling a comprehensive evaluation of textual, visual, and cross-modal inconsistencies. Each dataset provides a distinct annotation scheme, capturing various types of misinformation, from textual distortions to manipulated images and multimodal inconsistencies.

MMFakeBench Dataset. This dataset serves as a benchmark for multimodal misinformation detection. It categorizes misinformation into three primary types:

- **Textual Veracity Distortion (TVD)**: Fake or misleading textual claims.
- **Visual Veracity Distortion (VVD)**: Manipulated or AI-generated images.
- **Cross-Modal Consistency Distortion (CMM)**: Mismatches between text and images.

Each sample in MMFakeBench is annotated based on:

- Whether the claim text is factually correct.
- Whether the accompanying image has been manipulated.
- Whether the text-image pair is consistent or inconsistent.

The dataset provides a structured framework to evaluate misinformation detection across multiple manipulation types, incorporating diverse real-world scenarios.

DGM4 Dataset. The DGM4 (Grounding Multi-Modal Media Manipulation) dataset is a large-scale collection of manipulated and real news samples focusing on human-centric content. It contains approximately 230,000 samples, distributed as follows:

- **77,426** pristine (real) image-text pairs.
- **152,574** manipulated samples, generated using:
  - **Face Swap (FS)**: 66,722 samples.
  - **Face Attribute Manipulation (FA)**: 56,411 samples.
  - **Text Swap (TS)**: 43,546 samples.
  - **Text Attribute Manipulation (TA)**: 18,588 samples.
  - **Mixed Manipulation Pairs**: 32,693 samples combining text and image edits.

Factify Dataset. This dataset is a multimodal fact verification dataset containing 50,000 samples collected from Twitter and online news sources in the United States and India. Each sample consists of:

- **Claim text**: A short statement, often extracted from tweets.
- **Claim image**: The corresponding image that either supports or contradicts the claim.
- **OCR text**: Extracted text from the claim image.
- **Document text**: A news article serving as supporting evidence.
- **Document image**: An image from the referenced news article.

Samples in Factify are categorized into five classes:

- **Support_Text**: The claim text is supported by the document text, but images are dissimilar.
- **Support_Multimodal**: Both the claim text and image match the document text and image.
- **Insufficient_Text**: The document does not provide enough textual evidence to support or refute the claim.
- **Insufficient_Multimodal**: The document image matches the claim image, but the text lacks confirmation.
- **Refute**: The document contradicts both the claim text and the claim image.

The dataset provides a benchmark for multimodal fact verification, leveraging both textual and visual evidence to assess claim veracity.

Criteria	Definition
Critical Thinking and Skepticism	Challenges assumptions, probes deeper into claims, avoids taking information at face value.
Analytical Depth	Breaks down complex statements into verifiable components.
Systematic Approach	Follows a structured methodology in assessing sources, claims, and evidence.
Precision & Specificity	Clear, direct, and free from vague or overly broad wording.
Factual Accuracy	Focuses on verifying evidence, checking primary sources, and detecting misinformation.
Logical Consistency	Identifies contradictions, misleading narratives, or inconsistencies.
Source Credibility & Bias Detection	Evaluates the reliability of cited sources and potential biases.
Context Awareness	Considers the broader context surrounding the claim.
Comparative Thinking	Encourages cross-referencing with established facts or alternative perspectives.
Repeatability and Objectivity	Can be applied consistently across different cases without personal bias.

Table 4: Criteria for assessing the accuracy, credibility, and reliability of FCQs.

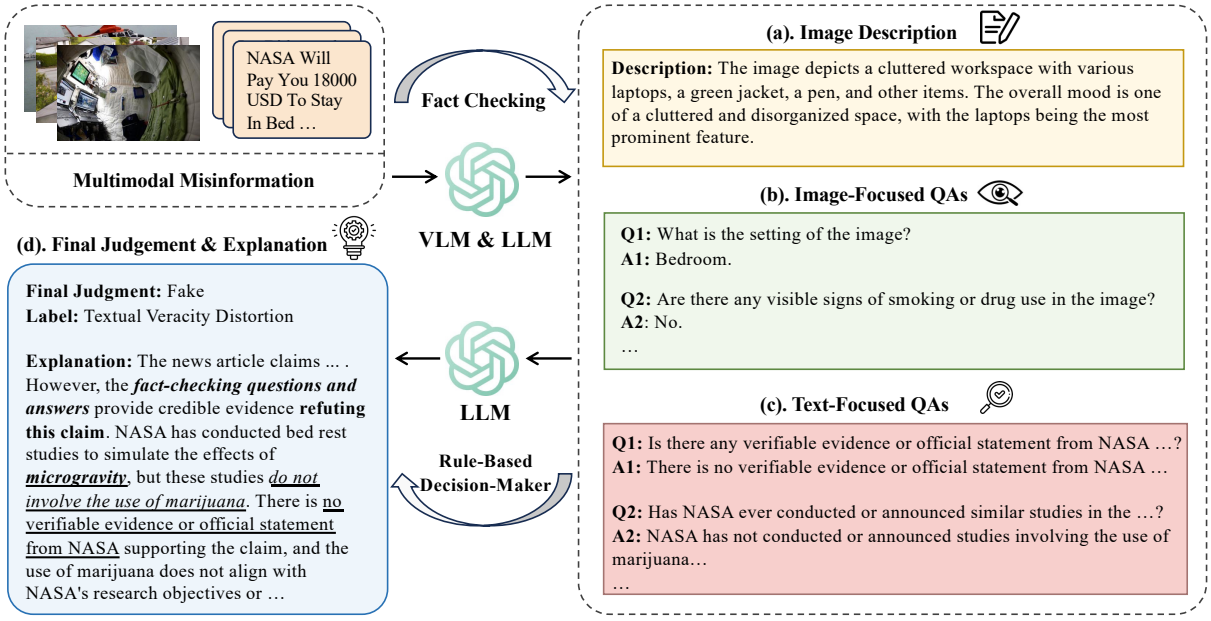

Figure 6: The overview pipeline of our LRQ-FACT framework consists of four key components: (a) Image Description, which provides detailed contextual descriptions of the image; (b) Visual FCQs, aimed at assessing the accuracy of the visual content; and (c) Textual FCQs, which detect textual inaccuracies, contradictions, or unsupported claims. Finally, all the gathered information is synthesized in (d) the Final Judgment & Explanation module, where a rule-based decision-maker generates both the prediction results and comprehensive explanations.

B Acknowledgment of AI Assistance in Writing and Revision

We utilized ChatGPT-4 for revising and enhancing sections of this paper.

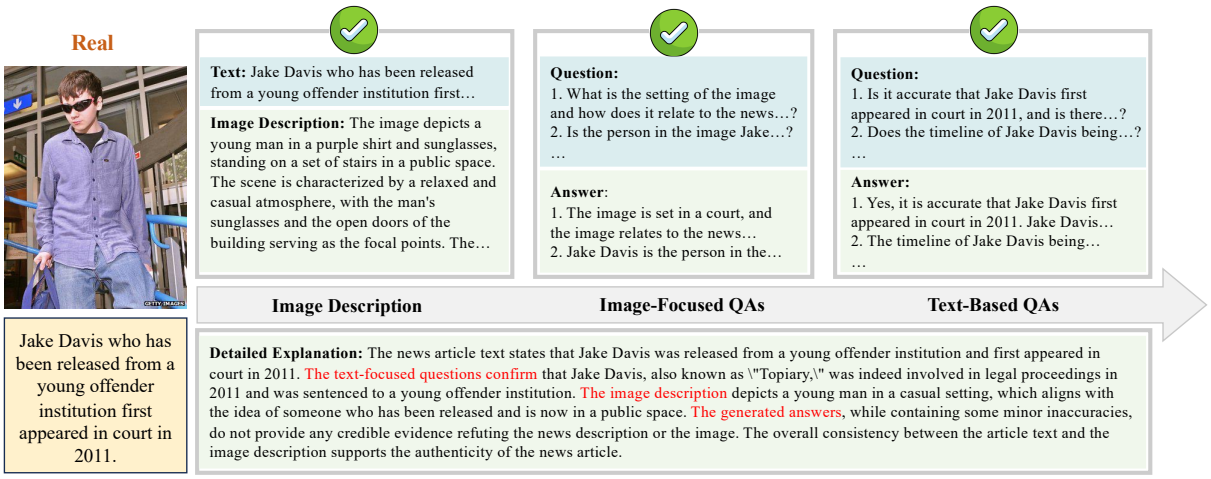

Figure 7: This example case illustrating the alignment between image and text in a fact-checking process. The generated questions verify key elements, ensuring consistency and accuracy in multimodal misinformation detection. This demonstrates how targeted questions and well-constructed descriptions enhance reliable fact-checking outcomes.

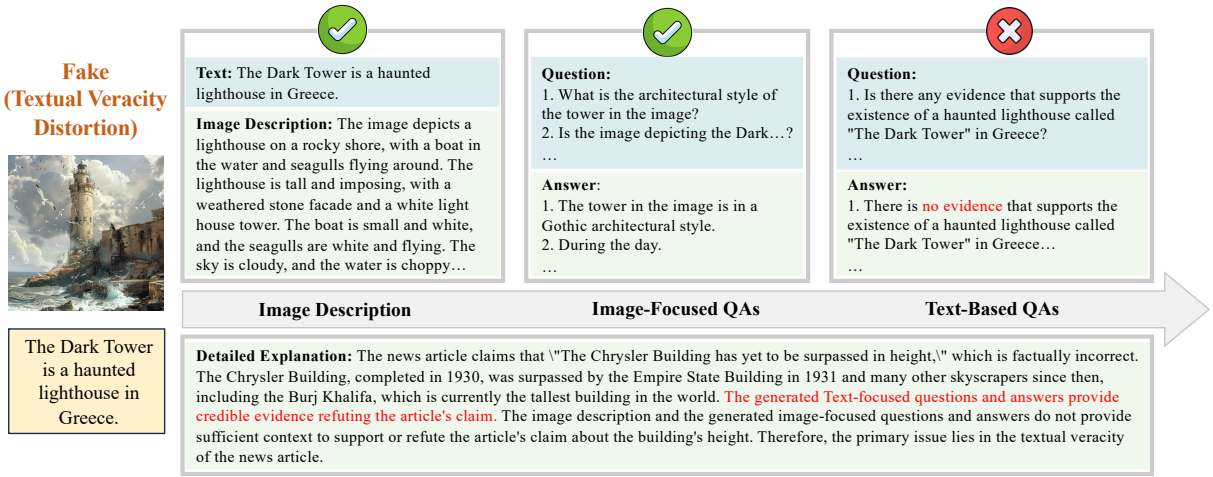

Figure 8: Example case illustrating textual veracity distortion. The image description aligns visually with the content, but fails to support the false textual claim about the haunted lighthouse's location in Greece. The generated questions are designed to detect inconsistencies, providing a thorough framework for fact-checking by scrutinizing both visual and textual elements.

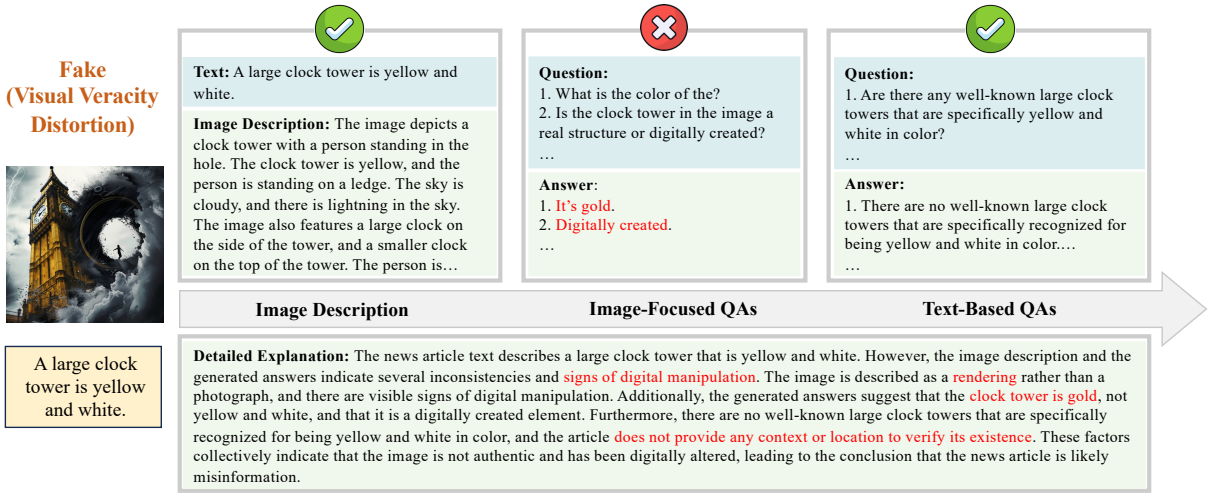

Figure 9: This example case highlighting visual manipulation. The description accurately conveys the textual claim about a yellow and white clock tower, but fails to reflect the digitally altered gold structure seen in the image. The questions focus on detecting visual discrepancies, such as the altered colors, and also probe the existence of such a clock tower, providing a reliable framework for evaluating both the image and text.

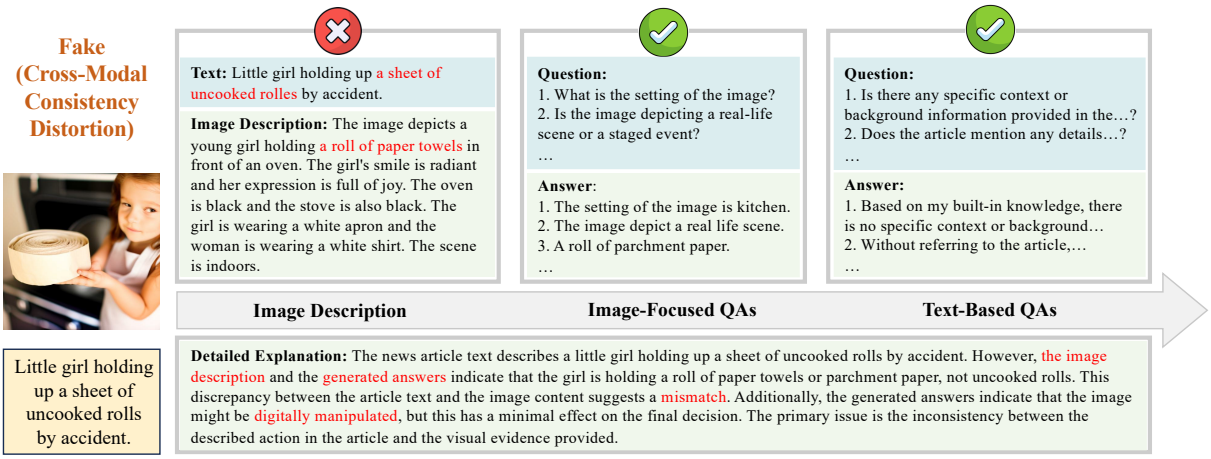

Figure 10: This example case demonstrating cross-modal distortion. The description is clear and helps visualize the article's scenario of a little girl holding uncooked rolls, while the image actually shows her holding paper towels. This mismatch between text and image points to a cross-modal distortion. The questions are well-crafted to reveal this inconsistency by focusing on both the scene and the object in the girl's hands, providing a reliable framework for identifying the discrepancy between the article and the image.

IMAGE DESCRIPTION PROMPT:

Please provide a detailed and comprehensive description of the image shown. Focus on identifying all visible elements including objects, people, setting, and any interactions or actions taking place. Describe the colors, textures, mood, and any other notable aspects that contribute to the overall context and significance of the image.

Figure 11: Structured prompt to generate detailed image descriptions.

VISUAL QUESTIONS PROMPT:

Given the following news article [**news text**], generate up to [**number of questions**] questions that are directly based on the news article and are designed to explore visual elements that could be present in an image related to the article.

Instructions for Question Generation:

Focus on generating questions that are directly relevant to the news article and the visual elements that could be present in an image. The questions should examine visible interactions, settings, actions, text, symbols, and specific objects mentioned in the article. Additionally, include questions that assess the authenticity of the image, such as whether it could have been AI-generated or contains any unusual or suspicious elements.

Avoid the following in your Questions:

- Do not mention any names.
- Do not ask questions about identification .
- Do not ask about personal details.
- Do not ask compound questions in a single sentence.

Example Questions:

1. What event is depicted in this image?
2. How are the people in the image interacting?
*3. Is the person in the image performing **[action from article]**?*
4. What are the technical aspects or tools used to create this image?
5. What emotions does this image evoke?
6. What are the main objects or elements visible in this image?
7. What unusual elements in the image might suggest digital manipulation or artificial creation?
...

Questions: *1. , 2. , ...*

Figure 12: Structured prompt to generate relevant visual questions.

VISUAL ANSWERS PROMPT:

You are an advanced AI model with access to a vast repository of knowledge and the capability of answering image questions. Your task is to answer the following questions [**generated questions**] based on the image [**image**]. While a news article [**news text**] is provided for context, you must answer the questions solely based on the image and not refer to the article's content.

Instructions for Answer Generation:

- Provide accurate, clear, and concise answers to each question.
- Your responses should be based entirely on the image.
- Do not reference or rely on the content of the provided news article when forming your answers.
- Each answer should be directly relevant to the question asked.

Avoid the following in your Answers:

- Provide accurate, clear, and concise answers to each question.
- Your responses should be based entirely on the image.
- Do not reference or rely on the content of the provided news article when forming your answers.
- Each answer should be directly relevant to the question asked.

Answers: *1. , 2. , ...*

Figure 13: Structured prompt to generate answers for the visual questions.

TEXTUAL QUESTIONS PROMPT:

Given the following news article [**news text**], analyze the text and formulate up to [**number of questions**] questions that probe the accuracy and verifiability of the information contained in the article. These questions should be designed to identify potential inaccuracies or areas that can be confirmed or challenged based on general knowledge or the text itself.

Instructions for Question Generation:

Focus on generating high-quality, fact-checking questions that can be answered directly through general knowledge that an LLM might possess. Identify and question significant factual claims, examine dates, locations, names, and other data mentioned in the article, and challenge any assumptions. The goal is to produce questions that facilitate direct verification of the facts stated in the article.

Aim to Generate:

- Questions that challenge the accuracy of specific claims made in the article and can be answered based on general knowledge.
- Questions that explore potential inconsistencies or contradictions within the article's content.
- Questions that assess the logical coherence and factual basis of the article's claims.

Avoid asking for:

- Information requiring external sources or verification beyond general knowledge.
- Speculative or opinion-based questions.

Example Questions:

1. Does the description of the "meeting between world leaders on March 5th" align with the known schedule of diplomatic events for that time?
2. Is the account of "a large protest taking place in front of City Hall" consistent with known reports of protests in that area during the stated period?
3. Does the timeline of "economic sanctions being imposed after the incident" logically follow the typical process for such actions?
4. Are the historical events referenced, such as "the financial crisis of 2008", accurately portrayed in the article?
...

Questions: *1. , 2. , ...*

Figure 14: Structured prompt to generate relevant textual questions.

TEXTUAL ANSWERS (W/O EVIDENCE) PROMPT:

You are an advanced AI model with access to a vast repository of knowledge. Your task is to answer the following questions [**generated questions**] based on your built-in knowledge. While a news article [**news text**] is provided for context, you must answer the questions solely based on your own knowledge and not refer to the article's content.

Instructions for Answering:

Provide accurate, clear, and concise answers to each question. Your responses should be based entirely on your general knowledge and the information you have learned. Do not reference or rely on the content of the provided news article when forming your answers. Each answer should be factually correct and directly relevant to the question asked.

Answers: *1. , 2. , ...*

Figure 15: Structured prompt to generate answers based on llm-knowledge for the relevant textual questions.

TEXTUAL ANSWERS (W/EVIDENCE) PROMPT:

You are an advanced AI tasked with evaluating the authenticity of a news article. Your task is to answer the following questions [**generated questions**] based on the provided factual document [**evidence**]. While a news article [**news text**] is provided for context, you must answer the questions solely based on the provided factual document and not refer to the article's content.

Instructions for Answering:

Provide accurate, clear, and concise answers to each question. Your responses should be based entirely on the provided factual document. If there was no factual answer for the question use your built-in knowledge to answer the question. Do not reference or rely on the content of the provided news article when forming your answers. Each answer should be directly relevant to the question asked.

Answers: *1. , 2. , ...*

Figure 16: Structured prompt to generate answers based on factual evidence for the relevant textual questions.

QUESTIONS QUALITY PROMPT:

You are an advanced AI tasked with evaluating the authenticity of a news article and its accompanying image. Your objective is to determine whether the provided questions [**generated questions**] effectively assess the accuracy, credibility, and reliability of the news article text.

Instructions:

Assess each question based on the following expert fact-checking criteria:
1. Critical Thinking and Skepticism: Does the question challenge assumptions, probe deeper into claims, and avoid taking information at face value?
2. Analytical Depth: Does it break down complex statements into verifiable components?
3. Systematic Approach: Does it follow a structured methodology in assessing sources, claims, and evidence?
4. Precision & Specificity: Is it clear, direct, and free from vague or overly broad wording?
5. Factual Accuracy: Does it focus on verifying evidence, checking primary sources, and detecting misinformation?
6. Logical Consistency: Does it help identify contradictions, misleading narratives, or inconsistencies?
7. Source Credibility & Bias Detection: Does it evaluate the reliability of cited sources and potential biases?
8. Context Awareness: Does it consider the broader context surrounding the claim?
9. Comparative Thinking: Does it encourage cross-referencing with established facts or alternative perspectives?
10. Repeatability & Objectivity: Can the question be applied consistently across different cases without personal bias?

Rating Scale:

- Relevant: The question is precise, well-structured, and effectively assesses factual accuracy, credibility, and logical consistency.
- Irrelevant: The question is vague, lacks depth, or fails to critically probe the credibility and factuality.

Answers:

Q1: [Relevant or Irrelevant]
Q2: [Relevant or Irrelevant]
...

Figure 17: Structured prompt to evaluate the quality of generated questions.

RULE-BASED DECISION-MAKER PROMPT:

Your objective is to determine whether the article and image are real or fake by analyzing the following information:

*1. News Article Text: **[news text]***

*2. Image Description: **[image description]***

Note: This description helps verify consistency with the news text. It is generally reliable but may contain minor discrepancies, such as using different terms like "ocean" instead of "water".

*3. Generated Visual Questions and Answers: **[generated visual FCQs]***

Note: These answers were generated by an AI and may contain mistakes, such as incorrect details regarding locations, names, dates, or objects. They might also incorrectly suggest that the image has been manipulated or is AI-generated. If the answers suggest manipulation or that the image is AI-generated, this should have very low effect on your final decision, especially if the image description and news article text do not contain such indications.

*4. Generated Textual Questions and Answers: **[generated textual FCQs]***

Note: These are based on the knowledge of GPT4-O, which is generally reliable but prone to hallucinations or contradictions with other provided information.

Instructions:

To make an accurate judgment of the multimodal misinformation, please follow these steps:

***Step 1.** Is there any credible objective evidence refuting the news description? If yes, assign the label: Textual Veracity Distortion. If no, continue to Step 2.*

***Step 2.** Is there any credible objective evidence refuting the news image? If yes, assign the label: Visual Veracity Distortion. If no, continue to Step 3.*

***Step 3.** Does the news caption match the content of the news image? If no, assign the label: Mismatch. If yes, and none of the above applies, assign the label: Real.*

Additional Guidelines:

1. Assess Overall Consistency: ...

2. Examine Details: ...

3. Analyze Facial Expressions and Body Language: ...

4. Identify Unrealistic Elements: ...

5. Cross-Modal Consistency: ...

6. Final Judgment: ...

7. Select the Most Relevant Label: ...

8. Provide a Detailed Explanation: ...

Example Output:

1. Final Judgment: *Fake*
2. Label: *Visual Veracity Distortion*
3. Explanation: *The image description mentions a "cat with pink eyes", which is highly unnatural and suggests the image is AI-generated. Additionally, ...*

---

1. Final Judgment: [Real or Fake]
2. Label: [Select one: Textual Veracity Distortion, Visual Veracity Distortion, Mismatch, Real]
3. Explanation: [Provide your explanation here]

Figure 18: Structured prompt to make the final decisions and provide an explanation.