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 responses. 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.

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.

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.

B Acknowledgment of AI Assistance in Writing and Revision

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

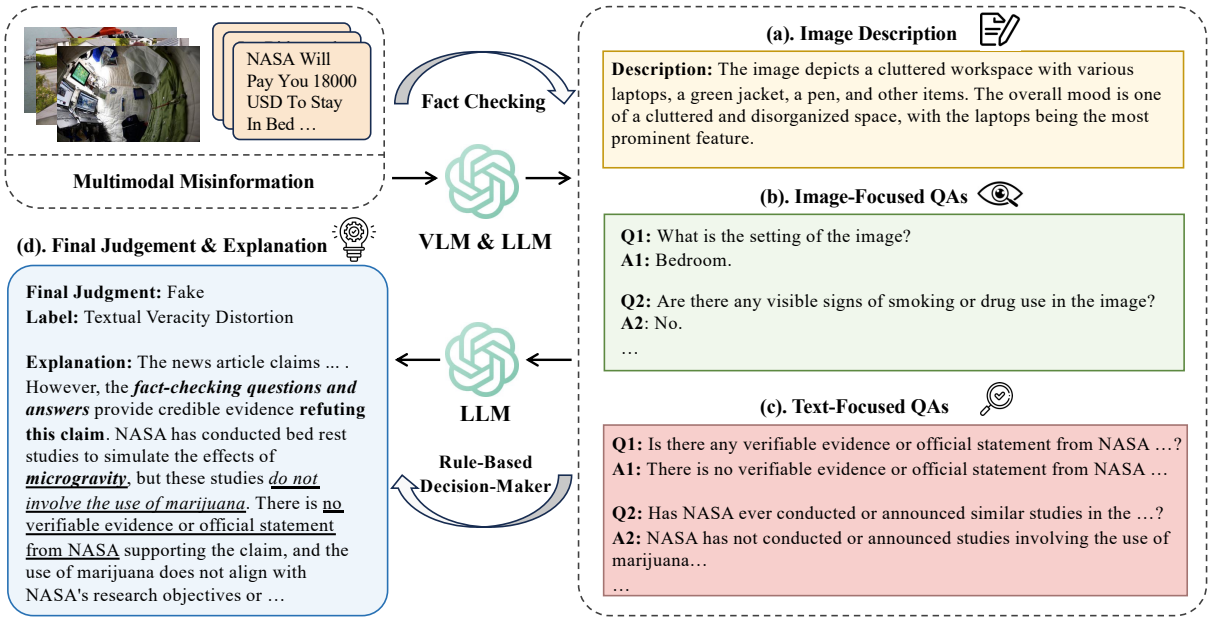

Figure 6: The overview pipeline of our LRQ-FACT framework consists of four key components: (a) Image Description, which provides detailed contextual descriptions of the image; (b) Visual FCQs, aimed at assessing the accuracy of the visual content; and (c) Textual FCQs, which detect textual inaccuracies, 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.

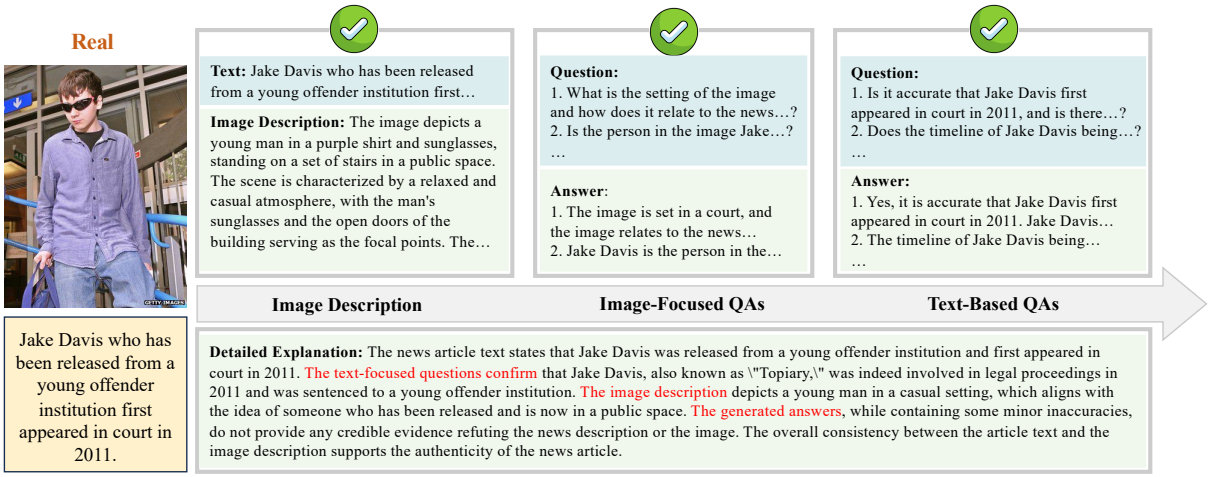

Figure 7: This example case illustrating the alignment between image and text in a fact-checking process. The generated questions verify key elements, ensuring consistency and accuracy in multimodal misinformation detection. This demonstrates how targeted questions and well-constructed descriptions enhance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

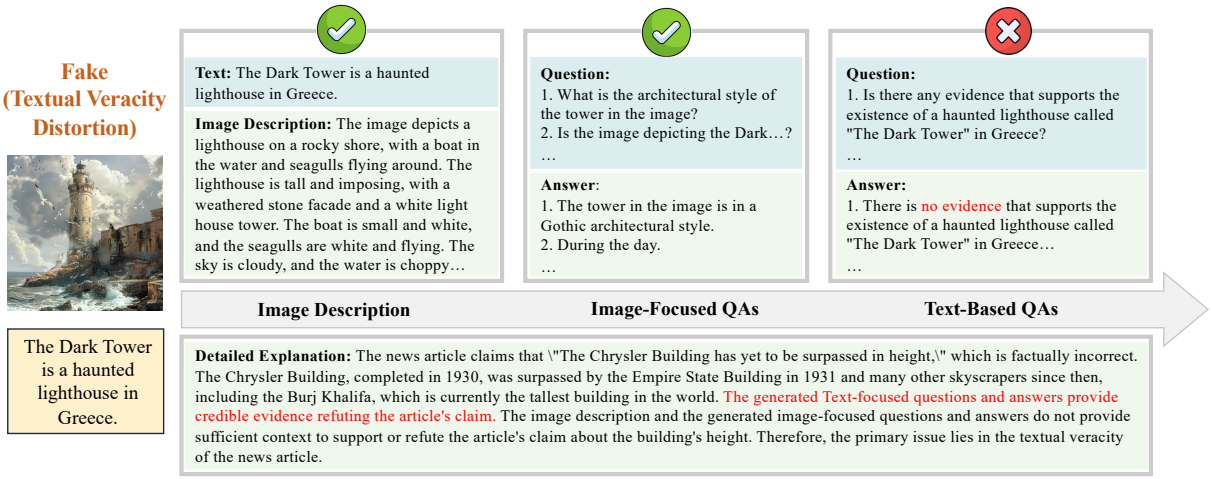

Figure 8: Example case illustrating textual veracity distortion. The image description aligns visually with the content, but fails to support the false textual claim about the haunted lighthouse's location in Greece. The generated questions are designed to detect inconsistencies, providing a thorough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

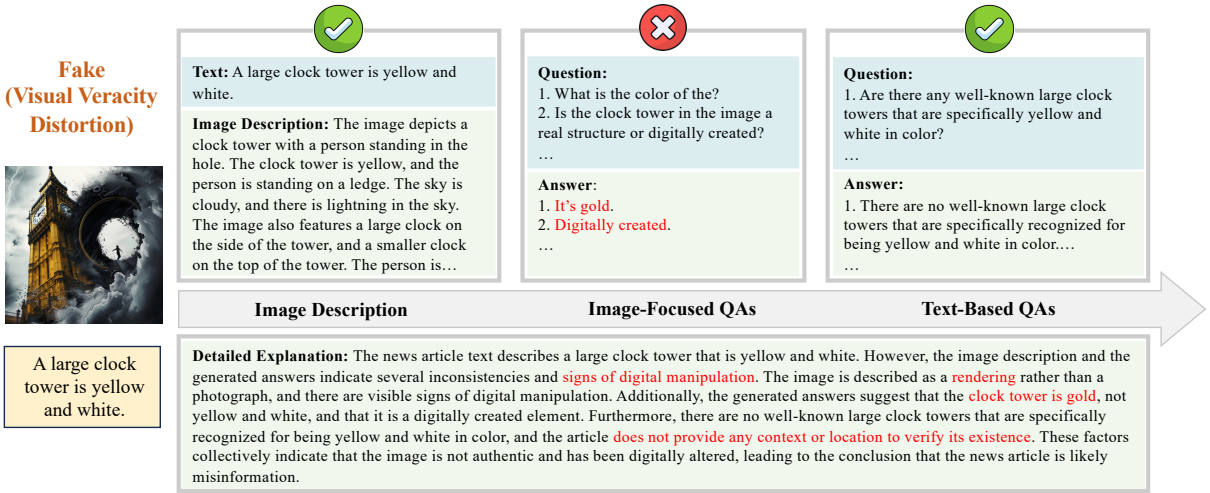

Figure 9: This example case highlighting visual manipulation. The description accurately conveys the textual claim about a yellow and white clock tower, but fails to reflect the digitally altered gold structure seen in the image. The questions focus on detecting visual discrepancies, such as the altered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

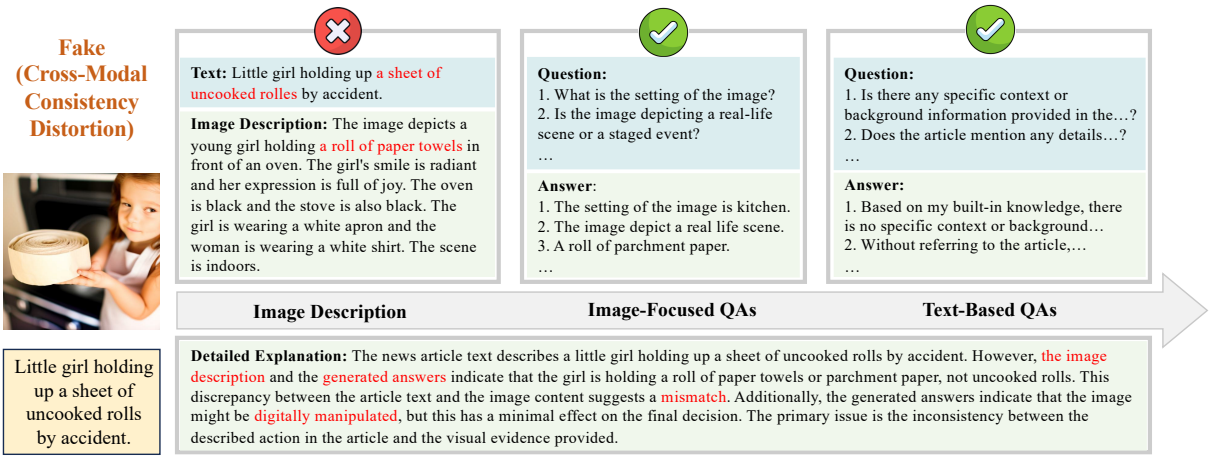

Figure 10: This example case demonstrating cross-modal distortion. The description is clear and helps visualize the article's scenario of a little girl holding uncooked rolls, while the image actually shows her holding paper towels. This mismatch between text and image points to a cross-modal distortion. 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.