Appendix

.1. Dataset

Table S1 and Table S2 present the prompts used to filter samples in the RW-Post dataset, aiming to reduce label leakage from images and to exclude samples where only the text contributes to veracity assessment.

Prompt for Image Verification: Label-Watermark Detection

Role: You are an expert in image verification and multimodal misinformation analysis.

Task: Given an input image, extract any textual content via OCR and determine whether the image contains watermark-like labels commonly used by fact-checking organizations (e.g., "fake", "altered", "misleading", "satire", etc.).

Requirements:

1. **Extract OCR text** from the image as accurately as possible.
2. **Detect watermark-like keywords** indicating the image has been labeled or classified. *fake, false, altered, misleading, miscaptioned, scam, satire, outdated, unproven, mixture, correct attribution, mostly false, mostly true, misattributed, composite image, no evidence, April Fools' Day*.
3. **Provide a concise explanation** for your decision.

Output Format:

```
{
  "watermarked": true/false,
  "matched_keywords": ["..."],
  "reason": ""
}
```

Table S1. Prompt for evaluating whether an image contains label-leaked watermarks

Prompt for Image Necessity Assessment

Role: You are a fact-checking expert.

Task: Determine whether the claim **requires an image** for verification.

Criteria for when an image is considered "required":

- The image would serve as direct evidence for the claim (e.g., confirming an event);
- The image would significantly influence users' judgment of the claim's truthfulness.

Output:

1. `yes` / `no` / `uncertain`
2. Brief explanation

Format:

```
{
  "image_required": "yes | no | unsure",
  "reason": ""
}
```

Table S2. Prompt for evaluating whether image context is required for a claim

.2. Prompts of Agents in AgentFact

Prompt for Strategy Planning Agent (Agent-SP)

Role: You are a fact-checking plan designer in a multi-agent fact-checking framework.

Task: Given the post content, claim, and context, generate or refine a verification plan that guides efficient and accurate fact checking. Apply appropriate fact-checking techniques (e.g., Divide and Conquer, Origin Tracing, Chain of Evidence, Cross-Verification, Temporal Consistency, Source Credibility, Logical Consistency).

Your Output Must Contain:

1. **Validation Logic** A concise, structured reasoning plan for analyzing the claim, indicating which techniques apply.
2. **Validation List (up to 3 items)** Original sentences from the post requiring verification. Each sentence must explicitly contain all essential information (no pronouns). Leave empty if none require direct validation.
3. **Search List (up to 3 items)** Key information that must be externally retrieved for fact checking. Ordered by priority and non-overlapping with the validation list.

Constraints:

- The validation list and search list must not overlap or contain redundant items.
- Include only information truly necessary for fact checking.
- Keep the plan concise but complete.

Output Format:

```
{
  "reasoning_steps": [ {"step": "", "method": "
      ", "details": ""}, ... ],
  "validation_list": [ {"sentence": "", "
      explanation": ""}, ... ],
  "search_list": [ {"information_needed": ""},
      ... ]
}
```

Prompt for Text Evidence Retrieval and Validation Agent (Agent-TR-1): Query Generation

Role: You are a fact-checking retrieval assistant responsible for generating high-quality search queries.

Task: Given an information need, claim and post context, produce a small set of SEO-effective, high-intent search queries suitable for Google. Queries should be specific, long-tail, and directly usable for evidence retrieval. Avoid redundancy with previous queries and ensure each query targets distinct information.

Guidelines:

- Generate focused, high-intent long-tail queries.
- Avoid duplicate or semantically similar queries.
- When necessary, break complex information into smaller searchable components.
- Tailor queries for reliability and relevance.
- Keep queries concise and avoid vague filler phrases.

Output Requirements:

- Generate at most one query per information item.
- Include only information worth retrieving externally.
- Queries must contain explicit keywords (no pronouns).

Output Format:

```
{ "queries": [ "best search query 1", "best
    search query 2" ] }
```

Prompt for Text Evidence Retrieval and Validation Agent (Agent-TR-2): Source Reliability Assessment

Role: You are an expert in digital literacy and online misinformation analysis. Your task is to assess the reliability and intent of a given website domain.

Task: Given a URL or domain, classify the source into one of the categories: `reliable`, `unreliable`, `satire`, `unsure`, `factcheck`.

Requirements:

1. **Identify the domain** (e.g., `cnn.com`, `theonion.com`).
2. **Evaluate source characteristics**, including whether it is:
   - listed in misinformation / disinformation databases;
   - frequently debunked by reputable fact-checkers;
   - known satire or parody;
   - legitimate, professional journalism;
   - a fact-checking organization.
3. **Provide a concise classification explanation** (2–4 sentences).
4. **Describe how a fact-checker should treat information** from this domain:
   - **Positive use**: information can generally be trusted;
   - **Reverse use**: presence of information is itself evidence of low credibility;
   - **Neutral/unsure**: requires strong corroboration.

Output Format:

```
{
  "source_identification": "",
  "type": "<reliable | unreliable | satire |
      unsure | factcheck>",
  "reasoning": "",
  "fact_checker_usage": ""
}
```

Prompt for Reasoning Agent (Agent-R)

Role: You are an expert fact-checking reasoning agent. Your task is to analyze the claim using structured reasoning steps and evidence with source-reliability judgments.

Task: Given the claim, post context, reasoning plan, retrieved text evidence (annotated with source reliability), and image-analysis results, execute each reasoning step in order and identify which evidence is relevant, irrelevant, or not required.

Requirements:

1. **Interpret the claim.** Produce a concise paraphrase capturing the core factual assertion. Output as `"my_understanding_of_claim"`.
2. **Follow the reasoning plan strictly.** Execute each step in the provided sequence.
3. **Evaluate evidence step-by-step.** For each reasoning step:
   - Identify relevant evidence and explain how it supports the analysis.
   - If no evidence is needed, state `"Evidence not required"`.
   - If no relevant evidence exists, state `"Relevant evidence not found"`.
   - You may optionally provide reliable evidence based on your own knowledge (with source, link, and reputation).
4. **Restrictions.**
   - The post itself cannot be used as evidence.
   - Conflicting evidence must be evaluated with respect to source reliability.
   - Absence of evidence does not imply falsity—assign confidence cautiously.
5. **Final confidence score (1–5).** Score reflects the sufficiency and reliability of evidence supporting the final assessment.

Output Format:

```
{
  "my_understanding_of_claim": "",
  "validation_result": {
    "reasoning_steps": [
      {
        "step_name": "",
        "description": "",
        "analysis_result": "",
        "relevant_evidence_summary": "",
        "relevant_text_evidence_list": [],
        "relevant_image_evidence_list": [],
        "evidence_based_on_my_knowledge": []
      }
    ],
    "direct_fact_check_evidence": {
      "analysis_result": "",
      "relevant_evidence_summary": "",
      "relevant_text_evidence_list": []
    },
    "final_sufficiency_confidence": ""
  }
}
```

Note: Keep reasoning concise but evidence-grounded.

Prompt for Image Retrieval and Analysis Agent (IR-1): Image Matching and Manipulation Detection

Role: You are an image comparison assistant supporting multimodal fact checking.

Task: Given a post image and a retrieved evidence image, analyze their visual relationship and assess whether the post image shows signs of manipulation.

Step 1: Classify Image Relationship Determine which of the following categories best describes the relationship:

- **Potentially From Same Source**: Nearly identical composition, contents, and configuration.
- **Same Event, Different Content**: Depict the same real-world event but differ in angle, timing, or framing.
- **No Close Relationship**: Unrelated subjects, events, or contexts.

Step 2: Manipulation Assessment If the relationship is not *No Close Relationship*, evaluate whether the post image shows signs of tampering based on:

- Self-analysis of the post image
- Direct comparison with the evidence image

Output Requirements: Provide a concise explanation for the relationship classification, estimate tampering probability (0–100), and give a short reasoning summary.

Output Format:

```
{
  "relationship": "",
  "relationship_reasoning": "",
  "tampering_probability": "",
  "tampering_reasoning": "",
  "confidence": ""
}
```

Notes:

- Focus on visually discriminative and fact-check–relevant features.
- Keep explanations clear and grounded in observable visual evidence.
- Leave tampering fields empty if the relationship is ”No Close Relationship”.

Prompt for Image Retrieval and Analysis Agent (IR-2): Image Miscaption Detection

Role: You are an image–text consistency analysis assistant for fact checking.

Task: Given a post image, its claim, and a text context of an evidence image, determine whether the post image is miscaptioned.

Step 1: Interpret the Claim Provide a concise paraphrase capturing the core factual assertion of the claim. Output as `"my_understanding_of_claim"`.

Step 2: Understand the Evidence Summarize the evidence image and text in your own words, including its event, context, or purpose. Assess alignment with the claim.

Step 3: Compare Temporal and Contextual Information Compare dates, locations, or actors if available and identify discrepancies.

Step 4: Miscaption Analysis

- comes from an unrelated event, place, or time;
- misrepresents who is involved or what is happening;
- substantially distorts the factual context.

Output Format:

```
{
  "my_understanding_of_claim": "",
  "Miscaption Rate": "",
  "Reasoning": ""
}
```

Scoring Guide

- 0–20: Image accurately supports the claim
- 30–50: Generally aligned but missing context
- 60–80: Provides a misleading impression
- 90–100: Unrelated or strongly contradicts the claim

Note: Focus strictly on factual alignment between the claim and image.

Prompt for Explanation Generation Agent (Agent-EG)

Role: You are the final reasoning and explanation agent in a fact-checking framework.

Task: Given the claim, post context, textual evidence, image-analysis results, and previous reasoning outputs, produce a final authenticity assessment with explanation and confidence.

Requirements:

1. **Interpret the claim.** Provide a concise paraphrase.
2. **Assess claim authenticity.**
   - **Coarse label**: TRUE / FALSE
   - **Fine-grained label**: TRUE / FALSE / UNPROVEN
3. **Decision principles.**
   - Strong refuting evidence → FALSE
   - Strong supporting evidence → TRUE
   - Insufficient evidence → UNPROVEN
4. **Reasoning and evidence citation.** Cite text and image evidence IDs.
5. **Confidence score (1–5).**

Output Format:

```
{
  "my_understanding_of_claim": "",
  "validation_result": {
    "2-class_authenticity_label": "",
    "3-class_authenticity_label": "",
    "reasoning_logic": "",
    "key_points": [
      "1. ...",
      "2. ...",
      "3. ..."
    ]
  },
  "confidence_level": ""
}
```

Note:

- Focus on the truth of the claim itself.
- Emphasize factual alignment and evidence reliability.

.3. Experimental Details: Fact-Checking Domain Filtering

To mitigate potential *label leakage* during evidence retrieval—particularly when external search engines return fact-checking articles that explicitly state the veracity of the claim—we applied a domain filtering strategy.

We excluded any evidence items whose source URLs contained substrings associated with known fact-checking organizations or services. This ensures that the model cannot trivially infer labels by relying on already-verified ground truth statements from professional sources. The filtering was implemented by checking whether the domain name or URL of the retrieved evidence contains any of the following substrings, which are associated with known fact-checking organizations or services. For brevity, a representative list is shown below; the complete list is available in our code repository.

- snopes
- politifact
- factcheck
- truthorfiction
- hoax-slayer
- eadstories
- opensecrets
- fullfact
- checkyourfact
- realitycheck
- fact-check
- ...

This filter was applied consistently to all models that utilized external search (e.g., LEMMA, DEFAME, AgentFact) during evidence retrieval. By removing these high-authority verification sources, we ensure that the model's performance more accurately reflects its ability to reason over raw evidence, rather than memorize or match against curated labels.

.4. Human evaluation

Figure S1 shows the interface used by human evaluators to score the outputs of different models, including the ground-truth results. The results of the three models (including the ground-truth) are randomly ordered for each sample, and the annotators are blind to which model produced each result.

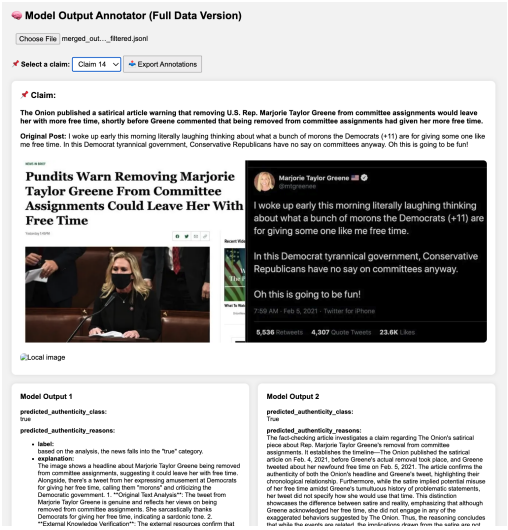

Figure S1. Screenshot of the model output annotator interface.

The following guideline was given to human evaluators, outlining the definitions of all evaluation metrics and the corresponding rating criteria.

Task Definition

The goal of the evaluator (the annotator) is to assess the outputs produced by fact verification models. Each model aims to determine the authenticity of a given claim and provide corresponding explanations.

Each claim originates from a social media post. The post text and image are provided as contextual information to assist understanding, but the final evaluation focuses strictly on the claim itself.

Evaluation Interface Overview

In the annotation interface, each claim is presented together with its textual content, image context, and outputs generated by three different models. Each model output is shown as an independent card containing:

- Authenticity judgment
- Reasoning and explanation
- Cited evidence

After finishing the evaluation for one claim, the results should be exported as a JSON file named according to the claim ID (e.g., `1.json`, `2.json`).

Each model output follows a structured JSON format:

```
{
  "predicted_authenticity_class": "True",
  "predicted_reasons": "...",
  "predicted_key_points": [...],
  "evidence_list": [...],
  "image_analysis_result": [...],
  "confidence_level": ""
}
```

Reasoning Hallucination

What to Evaluate This dimension evaluates the factual soundness of the model's reasoning. Specifically, check whether the reasoning provided in `predicted_authenticity_reasons` is directly supported by the evidence listed in `evidence_list`.

How to Judge

1. Read the reasoning and key points.
2. Identify factual claims such as numbers, dates, or causal statements.
3. Verify whether these claims are supported by the cited evidence.
4. If no concrete facts are invoked, evaluate logical coherence only.

Target Fields

- `predicted_authenticity_reasons`
- `predicted_authenticity_key_points`

Score	Criteria
none	Reasoning is coherent and well-supported by evidence.
mild	Minor unsupported assumptions without affecting the main conclusion.
severe	Core reasoning is speculative or disconnected from evidence.

Table S3. Scoring criteria for reasoning hallucination.

Evidence Usage Hallucination

What to Evaluate This dimension assesses whether the cited evidence is correctly used and accurately represented.

How to Judge

1. Examine each item in `evidence_list` and `image_analysis_result`.
2. Check whether the evidence supports the claims made.
3. Identify any fabrication, misinterpretation, or irrelevance.

Target Fields

- `evidence_list`
- `evidence_content_and_link`
- `image_analysis_result`

Score	Criteria
none	Evidence is accurate, relevant, and supports the reasoning.
partial	Evidence is partially misused or weakly relevant.
full	Evidence is fabricated or contradicts the conclusion.

Table S4. Scoring criteria for evidence usage hallucination.

Each criterion is rated on a three-level ordinal scale (0–2), mapped from {*none*, *mild/partial*, *severe/full*} to {0,1,2}.

Label Justification

What to Evaluate This dimension examines whether the predicted authenticity label is justified.

How to Judge

1. Identify the predicted label.
2. Compare it with the reasoning and evidence.
3. Check for overconfidence or contradiction.

Target Fields

- `predicted_authenticity_class`
- `predicted_authenticity_reasons`

Score	Criteria
justified	Label is consistent with reasoning and evidence.
overconfident	Label is stronger than supported by evidence.
hallucinated	Label contradicts the reasoning or evidence.

Table S5. Scoring criteria for label justification.