

Piecing It All Together: Verifying Multi-Hop Multimodal Claims

Haoran Wang[♡] Aman Rangapur[♡] Xiong Xiao Xu[♡] Yueqing Liang[♡]

Haroon Gharwi[♡] Carl Yang[♣] Kai Shu^{♣*}

[♡] Illinois Institute of Technology [♣] Emory University

{hwang219, arangapur, xxu85, yliang40, hgharwi}@hawk.iit.edu

{j.carlyang, kai.shu}@emory.edu

Abstract

Existing claim verification datasets often do not require systems to perform complex reasoning or effectively interpret multimodal evidence. To address this, we introduce a new task: multi-hop multimodal claim verification. This task challenges models to reason over multiple pieces of evidence from diverse sources, including text, images, and tables, and determine whether the combined multimodal evidence supports or refutes a given claim. To study this task, we construct MMCV, a large-scale dataset comprising 15k multi-hop claims paired with multimodal evidence, generated and refined using large language models, with additional input from human feedback. We show that MMCV is challenging even for the latest state-of-the-art multimodal large language models, especially as the number of reasoning hops increases. Additionally, we establish a human performance benchmark on a subset of MMCV. We hope this dataset and its evaluation task will encourage future research in multimodal multi-hop claim verification. Data and code are available: <https://mmcv-dataset.github.io/>

1 Introduction

Due to the rapid growth in AI-generated content (Huang et al., 2024a,b; Zhang et al., 2024; Jin et al., 2024b), it is difficult for automated fact-checking systems to keep up with verifying the accuracy of claims with multimodal evidence. This challenge is further exacerbated by the recent development of diffusion models such as DALL-E (Ramesh et al., 2021) and Stable Diffusion (Rombach et al., 2022), which can generate realistic images from textual prompts (Liu et al., 2024b). These powerful tools could enable attackers to produce misleading information (Wang and Shu, 2024; Pan et al., 2023c; Huang et al., 2024c; Gao et al., 2024; Jin et al., 2024a) at a low cost. Additionally, these claims

*Corresponding author

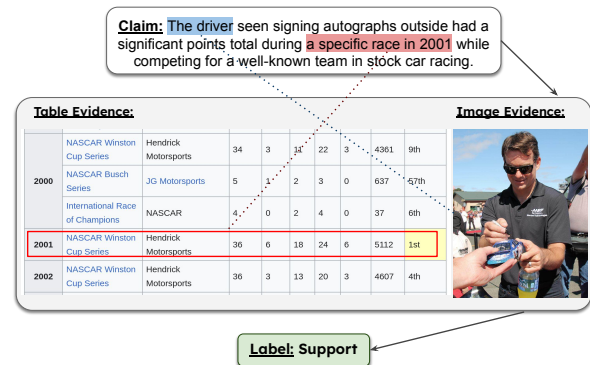


Figure 1: An illustration of a 2-hop claim from MMCV. To correctly verify this claim, the system must reason over both the image evidence and the table evidence.

often require multi-hop reasoning, where a set of connected evidence pieces leads to the final verdict of a claim (Yang et al., 2018). As a result, there is a need for automated tools to assist human fact-checkers in evaluating the veracity of multimodal multi-hop claims.

Claim verification, which involves assessing the veracity of an input claim against a collection of evidence, is a vital tool in combating the spread of misinformation (Thorne and Vlachos, 2018; Guo et al., 2022; Jin et al., 2022, 2023; Yang et al., 2022). However, verifying multi-hop multimodal claims introduces new challenges in both dataset construction and effective modeling. Unlike single-hop claims, which require only straightforward one-step reasoning, multi-hop claims require multiple reasoning steps to reach a final verdict. Furthermore, the inclusion of multimodal evidence requires models to understand and integrate information across various modalities, such as text, images, and tables, making it more complex to comprehend and extract relevant information. For instance, to verify the claim shown in Figure 1, a system must understand the semantic content of the image, integrate all relevant information from the table evidence, and apply multi-step reasoning to arrive at

Dataset	Multimodal	Multi-hop	Evidence Retrieval	Annotated Evidence	Annotated Label
FEVER (Thorne et al., 2018)	✗	✗	✓	✓	✓
Liar (Wang, 2017)	✗	✗	✗	✓	✓
FakeNewsNet (Shu et al., 2020)	✓	✗	✓	✗	✓
NewsCLIPpings (Luo et al., 2021)	✓	✗	✓	✗	✓
Factify (Mishra et al., 2022)	✓	✗	✗	✗	✗
COSMOS (Aneja et al., 2021)	✓	✗	✓	✗	✓
InfoSurgeon (Fung et al., 2021)	✓	✗	✓	✗	✓
Fauxtography (Zlatkova et al., 2019)	✓	✗	✗	✗	✓
HoVer (Jiang et al., 2020)	✗	✓	✓	✓	✓
MoCheg (Yao et al., 2023)	✓	✗	✓	✓	✓
MMCV (Ours)	✓	✓	✓	✓	✓

Table 1: Comparison between MMCV and other claim verification datasets. The columns indicate whether the dataset requires multimodal content, multi-hop reasoning, explanation generation, and whether it contains annotated evidence.

the final conclusion.

In this paper, we introduce the task of multi-hop multimodal claim verification to evaluate the veracity of multi-hop claims against multimodal evidence. To study this task, we construct **Multi-hop Multimodal Claim-Verification (MMCV)**, a dataset of 15K multi-hop claims paired with multimodal evidence that either SUPPORT or REFUTE each claim. To create the dataset, we develop a novel pipeline that uses large language models (LLMs) for data annotation, supported by human feedback. This method significantly reduces the workload on human annotators and cuts costs, while ensuring high quality and factual accuracy of the dataset. Our pipeline first uses LLMs to reformulate multi-hop multimodal question-answer pairs into atomic multi-hop claims and generate a set of candidate claims. These candidate claims are then modified to include additional hops and refined for fluency and clarity according to a set of annotation guidelines. To ensure the accuracy of the claims, we use a Retrieval-Augmented Generation (RAG)-based validation method to verify their validity. Finally, we ask a group of human annotators to score the claims based on their fluency, correctness, and clearness, and manually rewrite the claims that are below a certain threshold.

We establish performance baselines on MMCV using three state-of-the-art multimodal large language models (MLLMs) and highlight their limitations in verifying complex multimodal claims. We further demonstrate the challenges posed by the dataset, especially as the number of reasoning hops increases, by illustrating the constrained performance of various prompt techniques designed to

enhance MLLMs’ reasoning capabilities, including chain-of-thought, self-ask, and symbolic-guided reasoning. Additionally, we establish a human performance benchmark on a subset of MMCV.

Overall, we introduce a challenging multi-hop multimodal claim verification dataset that includes claims with up to 4 reasoning hops. These complex claims often consist of multiple sentences linked by coreference and demand evidence from various modalities, such as text, images, and tables. Table 1 provides a comparison between MMCV and existing popular claim verification datasets. While current datasets typically focus on either multimodal claims or multi-hop textual claims, none of them incorporate multi-hop multimodal claims that necessitate cross-modal reasoning. We hope that the introduction of MMCV and its corresponding evaluation task will inspire further research in complex multi-hop multimodal reasoning for claim verification. In summary, our contributions include:

- We introduce and formalize the multi-hop multimodal claim verification task.
- We develop a novel pipeline that leverages LLMs for data annotation, enhanced by human feedback, to construct a benchmark dataset for multi-hop multimodal claim verification. This method significantly lowers the cost and labor required to produce a large-scale dataset.
- We establish baseline performance on this task using MLLMs and human evaluation. Our analysis shows that this is a non-trivial task,

with several challenges that remain to be addressed in future work.

2 Background

Multimodal Claim Verification. Previous research on claim verification has primarily focused on textual data. However, with the growing recognition that misinformation often appears across multiple modalities and that multimodal misinformation is perceived as more credible and spreads faster than text-only misinformation, recent efforts have shifted toward verifying multimodal claims (Akhtar et al., 2023). As a result, several multimodal claim verification datasets have been proposed including FakeNewsNet (Shu et al., 2020), COSMOS (Aneja et al., 2021), InfoSurgeon (Fung et al., 2021), Factify (Mishra et al., 2022), Fauxtography (Zlatkova et al., 2019), and Moheg (Yao et al., 2023). However, to the best of our knowledge, there are no existing datasets for multi-hop multimodal claim verification, which challenges the system’s reasoning capability by requiring it to integrate and interpret multiple pieces of evidence from different modalities.

Multi-hop Reasoning. Verifying complex claims often requires multi-step (multi-hop) reasoning (Mavi et al., 2022), which requires combining information from multiple pieces of evidence to predict the veracity of a claim. Many recently proposed datasets are created to challenge a model’s ability to reason across multiple sentences or documents. These include MultiRC (Khashabi et al., 2018), QAngaroo (Welbl et al., 2018), ComplexWebQuestion (Talmor and Berant, 2018), HotpotQA (Yang et al., 2018), and HoVer (Jiang et al., 2020). In contrast to these datasets, MMCV incorporates context from various modalities, such as images and tables, further challenging the system’s ability to understand and integrate evidence from different sources.

Construct Synthetic Dataset with LLMs. The emergence of advanced large language models has sparked growing interest in automating the data annotation process using LLMs (Tan et al., 2024; Wu et al., 2024; Bao et al., 2024; Chen et al., 2024), driven by their advanced capabilities, including in-context learning (Dong et al., 2022) and learning from human feedback (Ouyang et al., 2022). (Wang et al., 2023) propose an explain-

then-generate pipeline using LLMs for iterative data synthesis, while (Pace et al., 2024) combine the Best-of-N and Worst-of-N sampling strategies to introduce the West-of-N approach. With this same objective, the multi-hop claims in MMCV are created and refined by LLMs using human feedback, following guidelines and rules specifically designed to enforce a multi-hop structure within each claim.

3 The MMCV dataset

The main goal of our work is to compile a diverse and extensive collection of multi-hop claims that require joint reasoning across evidence from different modalities, such as text, tables, and images, for verification. One approach to achieving this is to transform multimodal question-answering pairs into atomic claims and refine them to incorporate additional reasoning steps, making them more natural. However, there are two major challenges in creating such a dataset: first, *building a large-scale dataset is labor-intensive and costly*; second, in our pilot studies, we found that simply providing instructions to crowd workers and asking them to rewrite multi-hop claims is counterproductive, as *it is difficult to control quality and challenging for workers to create meaningful multi-hop claims*. Instead, we develop a pipeline that leverages the emerging capabilities of large language models to generate text and learn from feedback, with human input to ensure the quality of the final output.

In this approach, LLMs handle the mundane task of rewriting claims consistently according to the instructions, while human effort is significantly reduced to quality control of the final claims based on a set of guidelines. Figure 2 shows the overall workflow of our data construction pipeline, which contains three stages: LLM-Based Claim Generation (§3.1), LLM-Generated Claim Refinement (§3.2) and Claim Annotation by Human (§3.3).

3.1 Claim Generation

In this stage, we leverage the in-context learning capabilities of large language models to transform question-answer pairs from the MultimodalQA dataset (Talmor et al., 2021) into verifiable claims. To minimize the impact of in-context examples on the quality of the generated claims, we carefully craft a pool of 20 in-context examples and randomly select 3 for use during execution. The

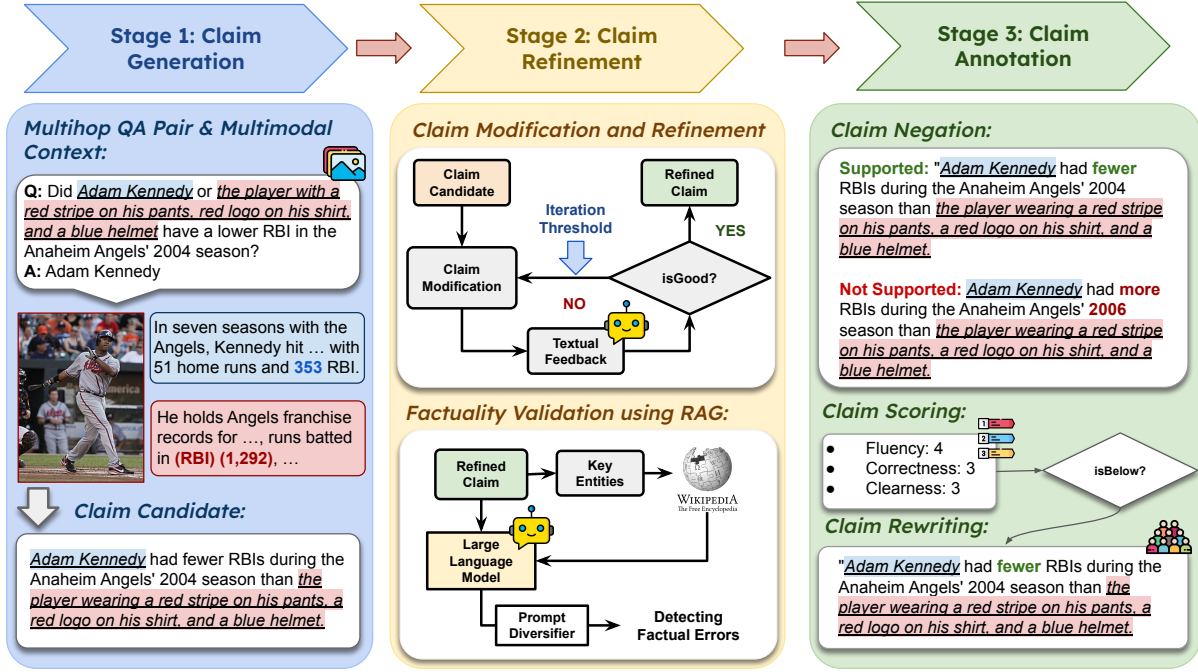


Figure 2: Overview of data collection flow chart for MMCV. In the first stage, we re-formulate question-answer pairs from *MultimodalQA* to generate candidate claims. In the second stage, we modify and refine the candidate claims, and apply a Retrieval-Augmented Generation (RAG)-based method to verify their correctness. In the final stage, we ask human annotators to rank the candidate claims to select the best one and label the final claims accordingly.

claims are formulated to ensure that no information is omitted from the original QA pairs and no new information is introduced. Since the claims are derived directly from the question and the correct answer, they are automatically labeled as SUPPORT. The prompt template for claim generation is listed in Appendix A.2.

3.2 Claim Refinement

After generating the initial claims from the question-answer pairs, we modify and refine them to ensure they are more naturally phrased and more accurately supported by the facts. Next, we review the claims for any factual errors that may have been introduced during the modification process and make corrections as needed.

Claim Modification and Refinement. To introduce additional reasoning steps to the claim candidate, we employ a modify-then-refine approach that iteratively enhances the quality of the modified claim candidate based on feedback from LLMs (Pan et al., 2023a). Specifically, we begin by identifying the Wikipedia entities mentioned in the answers from the question-answer pairs. If there is only one Wikipedia entity in the answer, we leave the claim candidate unchanged. However, if there are multiple Wikipedia entities, we use the sum-

maries of their respective Wikipedia articles as context and instruct the LLMs to modify the claim in such a way that it incorporates this contextual information to replace the entity, ensuring that the entity’s name does not appear directly in the claim.

To help LLMs understand the modification task, we provide them with 3-5 randomly selected in-context examples from a pool of hand-crafted examples. After modifying the claim, we obtain feedback from LLMs regarding the fluency, correctness, and clarity of the modified claim. The criteria used for this assessment are listed in the Appendix A.2. If the feedback suggests further improvement, the claim is sent back to the modification step, incorporating the LLMs’ feedback until a certain iteration threshold is reached. If the modified claim still does not pass the quality check, it is marked for manual review and revision by human annotators.

RAG-based Truthfulness Validation. Since we introduce additional contextual information from Wikipedia when modifying the claims, there is a risk that LLMs might hallucinate and produce outputs that are not faithful to the input context. To eliminate potential factual errors, we use a retrieval-augmented generation (RAG) (Lewis et al., 2020)-based pipeline to retrieve the full Wikipedia articles of the relevant entities and

validate the factual accuracy of the modified claims. To mitigate the impact of prompt sensitivity on the model’s output (Lu et al., 2022; Sclar et al., 2023), we diversify the prompts by randomly changing their format for each verification step. For instance, instead of consistently using `Is it true that {claim}?`, the prompt is randomly chosen from a set of equivalent alternatives, such as `Verify the following statement: {claim}` or `What evidence supports the claim that {claim}?`

3.3 Claim Annotation

At this stage, we have obtained claims that have been modified and refined by LLMs and factually validated by RAG-based pipelines. Next, we use LLMs to generate negated claims by applying a set of specific negation rules. We employ three distinct methods for generating these negated claims. For instance, given the claim, “*Since its construction in 1889, the Eiffel Tower in Paris attracts millions of visitors annually.*”, the results after applying the negation rules are as follows:

Negation

- ▷ **Word substitution:** *The Eiffel Tower in Paris houses millions of residents annually.*
- ▷ **Entity substitution:** *The Colosseum in Paris attracts millions of visitors annually.*
- ▷ **Temporal mutation:** *Ever since its construction in 2050, the Eiffel Tower has been Paris’s top tourist site.*

Next, a group of human annotators is tasked with evaluating the claims based on three dimensions: fluency, correctness, and clarity, scoring each dimension on a scale of 1 to 5. Fluency assesses how naturally the claim reads, as outputs generated by language models can sometimes sound artificial. Correctness evaluates whether the claim is factually accurate based on the evidence. Clarity determines if the claim is easily understood, as entity substitution might make it difficult to comprehend. Once the claims are scored, the average of the fluency, correctness, and clarity scores is calculated to determine the final score for each claim. If a claim’s final score falls below a predetermined threshold, it is flagged and sent back to the annotators for manual revision. Detailed annotation guidelines are listed in Appendix A.3.

4 Dataset Analysis

Dataset Statistics. MMCV contains 15,569 multi-hop multimodal claims, with their statistics

Data	1-hop	2-hop	3-hop	4-hop
# Claims	5,884	8,485	804	396
Ave. # Tokens in Claim	21.7	25.32	25.44	26.17
Max. # Tokens in Claim	48	58	51	63
# Text Evidence	2,590	7,323	1,142	760
# Image Evidence	1,979	2,948	634	512
# Table Evidence	1,315	6,699	636	312
# SUPPORT Labels	2,824	4,030	349	158
# REFUTE Labels	3,060	4,455	455	238

Table 2: Dataset Statistics of MMCV.

detailed in Table 2. The number of hops is determined by the count of multimodal evidence associated with each claim. The dataset includes a balanced distribution of SUPPORT and REFUTE claims. Specifically, there are 5,884 1-hop claims with an average of 21.7 tokens per claim; 8,485 2-hop claims averaging 25.32 tokens per claim; 804 3-hop claims with an average of 25.44 tokens per claim; and 396 4-hop claims averaging 26.17 tokens per claim. An example from the dataset is provided in Appendix A.1.

Multi-hop Reasoning Types. We provide examples of each reasoning type in Table 6. Most 1-hop and 2-hop claims require at least one supporting fact from either image or table evidence for verification. In contrast, the majority of 3-hop and 4-hop claims require evidence from all three modalities. The process of removing a bridge entity and replacing it with a relative clause or phrase significantly increases the informational load of a single hypothesis. As a result, some 3-hop and 4-hop claims are relatively longer and exhibit complex syntactic and reasoning structures. Our experimental results also indicate that the difficulty for models to verify claims escalates as the hop count increases.

5 Experiments and Results

In this section, we discuss our experiment settings (§5.1), the experiment results (§5.2), and the error analysis (§5.3). We begin by formally defining the MMCV task below.

Task Definition. The formulation of multi-hop multimodal claim verification is defined as follows: Given a claim C , and a list of multimodal evidence $\mathcal{E}(C)$, which includes text, images,

Retrieval	Model	1-hop			2-hop			3-hop			4-hop		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Closed-book	GPT-4o	76.86	72.94	71.79	67.96	63.30	60.66	62.88	58.89	56.17	67.93	62.39	61.20
	GEMINI	75.67	71.44	70.15	69.10	64.19	61.73	66.74	61.10	58.44	63.78	59.90	58.69
	LLAVA	64.18	63.78	63.57	64.06	63.93	63.87	66.78	66.81	66.76	64.64	64.84	64.64
Open-book	GPT-4o	76.95	72.95	71.78	68.03	63.24	60.53	62.67	58.78	56.08	67.75	62.46	61.35
	GEMINI	79.58	79.25	79.20	72.38	71.85	71.66	66.37	65.90	65.86	67.21	66.86	66.97
	LLAVA	62.86	59.68	57.21	64.17	62.48	61.50	65.47	64.64	63.76	66.50	66.76	66.42

Table 3: We report the Precision, Recall, and F1 scores of various MLLMs on MMCV for zero-shot multimodal claim verification. In the closed-book setting, the model verifies the claim without access to any external knowledge sources. In the open-book setting, the model is provided with a set of gold evidence. The best-performing model for each hop is highlighted in **Green** for both settings.

and tables, the system must reason over all the evidence and predict the label of the claim as either SUPPORT or REFUTE.

5.1 Experiment Settings

As there are no existing models specifically designed for multi-hop multimodal supervised claim verification, we conduct our experiments using MLLMs. Moreover, previous studies in textual claim verification and multimodal claim verification indicate that LLMs and MLLMs can significantly enhance task performance compared to traditional supervised approaches (Pan et al., 2023b; Wang and Shu, 2023; Li et al., 2024; Geng et al., 2024b). Furthermore, supervised methods often require extensive annotated corpora, which are difficult to acquire and limit domain transferability, as training data typically covers only a single domain.

Zero-shot Claim Verification. We establish performance baselines for zero-shot multimodal claim verification using various MLLMs under two settings. In the *closed-book setting*, the model does not retrieve information from external knowledge sources and must rely on its parametric (internal) knowledge to verify the claim. In the *open-book setting*, the model is provided with a set of gold evidence. Specifically, we use the prompt from (Geng et al., 2024b), which extracts the models’ predictions, explanations, and confidence levels. The prompt is listed in Appendix A.2. We use macro precision, recall, and F-1 score to evaluate the model performance.

MLLM. We utilize two state-of-the-art MLLMs: GPT-4o (Achiam et al., 2023) and Gemini 1.5 Flash (Team et al., 2023). Additionally, we

evaluate the performance of an open-source MLLM, LLaVA-V1.5-7B (Liu et al., 2024a), on MMCV. The temperature is set to 0.0, and the maximum number of tokens is set to 5000.

Prompts for Enhanced Reasoning In addition to the prompt mentioned above, we conduct experiments using specialized prompting techniques aimed at eliciting reasoning from LLMs, such as Chain-of-Thought (Wei et al., 2022) and Self-Ask (Press et al., 2023). We also test symbolic-guided reasoning prompts like ProgramFC (Pan et al., 2023b) and Visual Programming (Gupta and Kembhavi, 2023). To minimize the overall cost of the experiments, we randomly select 100 examples from each hop of the MMCV dataset for testing. The experiments are conducted using open-book setting.

Human Performance To benchmark human performance on our dataset, we used the same randomly selected examples employed in the enhanced reasoning prompt experiments. We recruited four experts in automated fact-checking research to classify multihop claims from MMCV based on the provided evidence. The SMART (Chew et al., 2019) framework¹ was used to deploy the annotation task, and human performance was evaluated using the macro F-1 score.

5.2 Experiment Results

Main Results. We report the comprehensive results of the three MLLMs on MMCV in Table 3, highlighting the best-performing models for each hop under both open-book and closed-book

¹<https://github.com/RTIInternational/SMART>

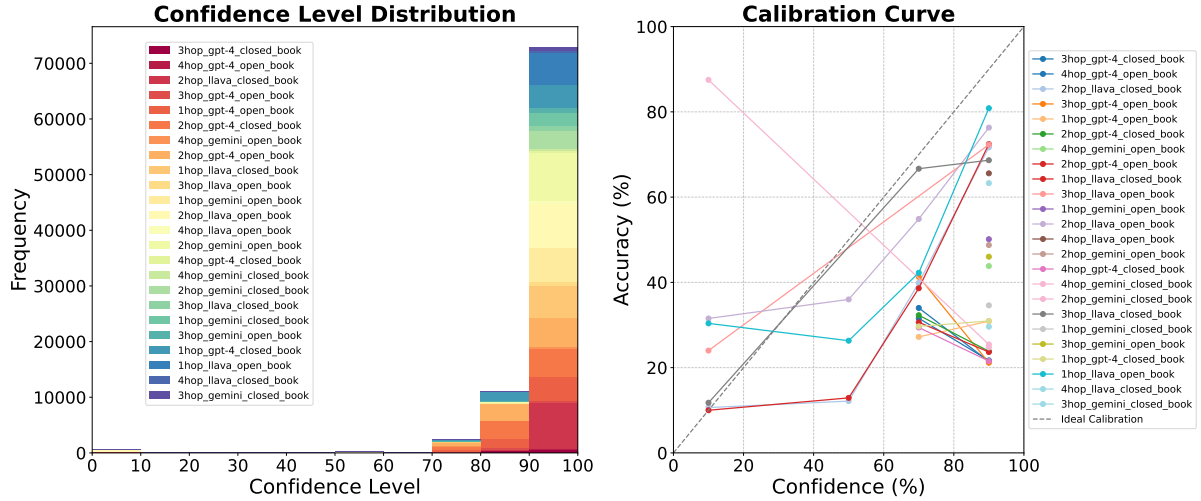


Figure 3: The left figure shows the confidence score distribution of GPT4-o, Gemini, and LLaVA on MMCV under both open-book and closed-book settings, categorized by the number of hops. The right figure shows their calibration curves.

settings. Overall, Gemini 1.5 outperforms others in the open-book setting with an average F-1 of 70.92, while LLaVA achieves the highest performance in the closed-book setting with an average F-1 of 66.77. This is surprising, given that LLaVA is a much smaller model compared to GPT4-o and Gemini, and therefore possesses less parametric knowledge. Upon manually analyzing a subset of 100 randomly selected outputs from LLaVA, we found that the model frequently hallucinates, even when it predicts the correct label, particularly as the hop count increases. This is consistent with its open-book performance, where its accuracy declines when provided with gold evidence. Additionally, we observe that GPT4-o performs slightly better in closed-book settings than in open-book settings, suggesting a tendency to hallucinate. In contrast, Gemini’s performance drops significantly in closed-book settings compared to open-book, demonstrating its robustness in effectively utilizing provided gold evidence.

Confidence Level Analysis The left panel of Figure 3 presents the confidence (Geng et al., 2024a) distributions for all three MLLMs, categorized by the number of hops and divided into 10 intervals. The results show that the majority of the MLLMs are concentrated in the 90-100 confidence range, with only a small number exhibiting low confidence (0-10 range), which occurs solely in open-book settings. This indicates that the MLLMs consider the provided gold evidence.

Model	Method	1-hop	2-hop	3-hop	4-hop
GEMINI 1.5	CoT	78.52	69.66	67.45	70.24
	Self-Ask	75.47	66.58	60.94	70.67
	Symbolic	74.89	63.82	54.61	72.36
GPT4-o	CoT	80.43	83.33	71.20	72.99
	Self-Ask	77.42	80.12	70.52	75.23
	Symbolic	80.56	78.78	68.72	75.67

Table 4: Results of Gemini and GPT4-o on 100 randomly sampled claims for each hop using three types of reasoning prompts. Model performance is evaluated using F-1 score.

The right panel of Figure 3 displays the calibration curves, illustrating the relationship between the models’ confidence levels and their actual classification accuracy. These curves reveal a positive correlation between confidence and accuracy for 1-hop and 2-hop claims, as exemplified by the red line (GPT-4-o on 2-hop), the teal line (LLaVA on 1-hop), and the purple line (Gemini on 1-hop). In contrast, the downward curves, mostly observed in 3-hop and 4-hop claims, suggest that the models tend to be overconfident when classifying more complex claims. Additionally, the results indicate that open-book settings generally have better-calibrated confidence scores than closed-book settings, further suggesting that the models exhibit overconfidence when not provided with gold evidence.

Annotator	# Hops			
	1-hop	2-hop	3-hop	4-hop
<i>Annotator 1</i>	83.33	86.20	78.42	79.82
<i>Annotator 2</i>	82.46	88.29	79.45	82.16
<i>Annotator 3</i>	80.60	90.53	80.62	85.24
<i>Annotator 4</i>	79.64	86.50	82.32	83.87

Table 5: Results of human performance on 200 random samples. Performance are measured by F-1 score.

Reasoning Prompt Results. Table 4 reports the performance of Gemini and GPT4-o on the randomly sampled subset of MMCV under open-book settings using various prompts that elicit LLMs’ reasoning abilities. For symbolic approach, we ask LLMs to first generate a Python-like program that decomposes the multi-hop claim into a set of function calls that describe the reasoning steps required to verify the claim, and use the symbolic information provided by the generated program to elicit better step-by-step reasoning from the model. We observe that GPT-4-o gains more from the enhanced reasoning prompt compared to Gemini, achieving a higher average F1 score of 75.93 in symbolic guided reasoning, whereas Gemini attains an average F1 score of 66.42 for the same task. Additionally, we found that Symbolic approach are more effective on 4-hop claims, having a higher F1 score than CoT and self-ask. However, this observation is different on simpler 2-hop and 3-hop claims, where CoT appears to be more effective.

Human Performance Results To establish human performance on our dataset, we randomly sampled 200 examples, with 50 examples from each hop from MMCV. We recruited four annotators to perform claim verification given the gold evidence. We trained our annotators on the task by providing them with guidelines and sample annotations to ensure consistency and accuracy in their evaluations. After training, the annotators independently verified each claim using the provided gold evidence, allowing us to assess the human baseline performance on the dataset. Table 5 reports the results from the human annotators. We observe that the human annotators achieve very high performance in verifying the claims across all 4 hops. The human performance is 23.3% and 27.3% higher than the best-performing MLLMs on 3-hop and 4-hop

claims respectively. This suggests that although MLLMs perform relatively well, there is still room for improvement to match human performance.

5.3 Error Analysis

Figure 5, 6, and 7 shows the error analysis of the false positive examples from GPT4-o, Gemini, and LLaVA respectively. We observe that visual misinterpretation is a major issue, with the system often misidentifying or miscontextualizing image elements. This problem is especially pronounced in examples involving sports logos and movie posters, highlighting the need for improvements in the visual processing component.

Another notable issue is the system’s handling of temporal and factual information. Errors related to player career timelines and historical events reveal shortcomings in temporal reasoning and the integration of world knowledge. The system’s confidence levels, often between 80% and 100% for incorrect predictions, suggest a miscalibration in certainty estimation. This overconfidence in erroneous conclusions highlights the need for a more refined approach to confidence scoring.

Last but not least, examples from higher hop categories reveal significant weaknesses in handling complex reasoning tasks. The system often struggles with multi-step logical inferences, frequently failing to coherently link disparate pieces of information. This limitation is especially problematic for claims that require advanced analysis or the cross-referencing of multiple facts.

6 Conclusion

In this paper, we introduce MMCV, a multi-hop multimodal claim verification dataset that requires models to aggregate information from up to four multimodal evidence to verify a claim. To create this large-scale dataset, we developed a novel data collection pipeline that leverages the capabilities of LLMs combined with human feedback. Specifically, our approach includes a module that iteratively refines modified claims using feedback from a judge LLM based on a set of predefined criteria, as well as an actuality validation module that employs RAG to ensure the factual accuracy of the claims. Our results show that state-of-the-art MLLMs struggle to verify more complex claims as the number of reasoning hops increases, often displaying overconfidence in their predictions. We also present findings from experiments utilizing

prompts tailored to enhance the reasoning abilities of MLLMs, alongside human performance benchmarks for comparison. Additionally, we categorize and provide a detailed error analysis of false positive results from each model. We hope that MMCV will inspire the development of models capable of conducting complex, multi-hop reasoning in the challenging task of multimodal claim verification.

7 Limitations

We identify two main limitations of MMCV. First, the construction of MMCV depends on in-context learning coupled with self-refinement to convert a natural language question-answer pair into a multi-hop claim. While this method has proven to be effective, it may face difficulties when dealing with questions with intricate grammar structures and logical structures. This arises from the difficulty in conveying complex grammatical rules to the language model through a limited number of demonstrations within a constrained context size. Second, our aggregation method purely relies on LLMs themselves, which could introduce potential hallucination problems. On the other hand, by using a more robust logic solver could help with the hallucination issues, but there would be a tradeoff between the applicability and the robustness of the model.

8 Ethical Statement

Biases. We acknowledge the possibility of biases existing within the data used for training the language models, as well as in certain factuality assessments. Unfortunately, these factors are beyond our control.

Intended Use and Misuse Potential. Our models have the potential to verify complex multimodal claims. However, it is essential to recognize that they may also be susceptible to misuse by malicious individuals. Therefore, we strongly urge researchers to approach their utilization with caution and prudence.

Environmental Impact. We want to highlight the environmental impact of using large language models, which demand substantial computational costs and rely on GPUs/TPUs for training, which contributes to global warming. However, it is worth noting that our approach does not train such models from scratch. Instead, we use few-shot in-context learning. Nevertheless, the large language models we used in this paper are likely running on GPU(s).

Acknowledgements

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 17STQAC00001-07-04, NSF awards (SaTC-2241068, IIS-2339198), a Cisco Research Award, and a Microsoft Accelerate Foundation Models Research Award. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security or the National Science Foundation. This research was partially supported by the National Institute Of Diabetes And Digestive And Kidney Diseases of the National Institutes of Health under Award Number K25DK135913.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, 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](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5430–5448, Singapore. Association for Computational Linguistics.
- Shivangi Aneja, Chris Bregler, and Matthias Nießner. 2021. Cosmos: Catching out-of-context misinformation with self-supervised learning. *arXiv preprint arXiv:2101.06278*.
- Han Bao, Yue Huang, Yanbo Wang, Jiayi Ye, Xiangqi Wang, Xiuyin Chen, Mohamed Elhoseiny, and Xiangliang Zhang. 2024. Autobench-v: Can large vision-language models benchmark themselves? *arXiv preprint arXiv:2410.21259*.
- Dongping Chen, Yue Huang, Siyuan Wu, Jingyu Tang, Liuyi Chen, Yilin Bai, Zhigang He, Chenlong Wang, Huichi Zhou, Yiqiang Li, et al. 2024. Gui-world: A dataset for gui-oriented multimodal llm-based agents. *arXiv preprint arXiv:2406.10819*.
- Rob Chew, Michael Wenger, Caroline Kery, Jason Nance, Keith Richards, Emily Hadley, and Peter Baumgartner. 2019. Smart: an open source data labeling platform for supervised learning. *Journal of Machine Learning Research*, 20(82):1–5.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

- Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi 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, Online. Association for Computational Linguistics.
- Chujie Gao, Qihui Zhang, Dongping Chen, Yue Huang, Siyuan Wu, Zhengyan Fu, Yao Wan, Xiangliang Zhang, and Lichao Sun. 2024. The best of both worlds: Toward an honest and helpful large language model. *arXiv preprint arXiv:2406.00380*.
- Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koepl, Preslav Nakov, and Iryna Gurevych. 2024a. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595.
- Jiahui Geng, Yova Kementchedjheva, Preslav Nakov, and Iryna Gurevych. 2024b. Multimodal large language models to support real-world fact-checking. *arXiv preprint arXiv:2403.03627*.
- Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. [A survey on automated fact-checking](#). *Transactions of the Association for Computational Linguistics*, 10:178–206.
- Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962.
- Baixiang Huang, Canyu Chen, and Kai Shu. 2024a. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *arXiv preprint arXiv:2408.08946*.
- Baixiang Huang, Canyu Chen, and Kai Shu. 2024b. Can large language models identify authorship? *arXiv preprint arXiv:2403.08213*.
- Yue Huang, Zhengqing Yuan, Yujun Zhou, Kehan Guo, Xiangqi Wang, Haomin Zhuang, Weixiang Sun, Lichao Sun, Jindong Wang, Yanfang Ye, et al. 2024c. Social science meets llms: How reliable are large language models in social simulations? *arXiv preprint arXiv:2410.23426*.
- Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. [HoVer: A dataset for many-hop fact extraction and claim verification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.
- Yiqiao Jin, Minje Choi, Gaurav Verma, Jindong Wang, and Srijan Kumar. 2024a. Mm-soc: Benchmarking multimodal large language models in social media platforms. *arXiv preprint arXiv:2402.14154*.
- Yiqiao Jin, Yeon-Chang Lee, Kartik Sharma, Meng Ye, Karan Sikka, Ajay Divakaran, and Srijan Kumar. 2023. Predicting information pathways across online communities. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1044–1056.
- Yiqiao Jin, Xiting Wang, Ruichao Yang, Yizhou Sun, Wei Wang, Hao Liao, and Xing Xie. 2022. Towards fine-grained reasoning for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5746–5754.
- Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024b. Agentreview: Exploring peer review dynamics with llm agents. *arXiv preprint arXiv:2406.12708*.
- Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. [Looking beyond the surface: A challenge set for reading comprehension over multiple sentences](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2024. [Self-checker: Plug-and-play modules for fact-checking with large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 163–181, Mexico City, Mexico. Association for Computational Linguistics.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. 2024b. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*.
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. [Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

- Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. [NewsCLIPPings: Automatic Generation of Out-of-Context Multimodal Media](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6817, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Vaibhav Mavi, Anubhav Jangra, and Adam Jatowt. 2022. A survey on multi-hop question answering and generation. *arXiv preprint arXiv:2204.09140*.
- 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*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Alizée Pace, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2024. West-of-n: Synthetic preference generation for improved reward modeling. *arXiv preprint arXiv:2401.12086*.
- Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023a. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*.
- Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023b. [Fact-checking complex claims with program-guided reasoning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.
- Yikang Pan, Liangming Pan, Wenhui Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023c. [On the risk of misinformation pollution with large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, Singapore. Association for Computational Linguistics.
- Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. [Measuring and narrowing the compositionality gap in language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.
- 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.
- Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. *arXiv preprint arXiv:1803.06643*.
- Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hananeh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*.
- Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansoor Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- James Thorne and Andreas Vlachos. 2018. [Automated fact checking: Task formulations, methods and future directions](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- Haoran Wang and Kai Shu. 2023. [Explainable claim verification via knowledge-grounded reasoning with large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6288–6304, Singapore. Association for Computational Linguistics.

- Haoran Wang and Kai Shu. 2024. [Trojan activation attack: Red-teaming large language models using steering vectors for safety-alignment](#). In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24*, page 2347–2357, New York, NY, USA. Association for Computing Machinery.
- Ruida Wang, Wangchunshu Zhou, and Mrinmaya Sachan. 2023. [Let’s synthesize step by step: Iterative dataset synthesis with large language models by extrapolating errors from small models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11817–11831, Singapore. Association for Computational Linguistics.
- William Yang Wang. 2017. [“liar, liar pants on fire”: A new benchmark dataset for fake news detection](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.
- Siyuan Wu, Yue Huang, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Xiangliang Zhang, Jianfeng Gao, Chaowei Xiao, et al. 2024. Unigen: A unified framework for textual dataset generation using large language models. *arXiv preprint arXiv:2406.18966*.
- Ruichao Yang, Xiting Wang, Yiqiao Jin, Chaozhuo Li, Jianxun Lian, and Xing Xie. 2022. Reinforcement subgraph reasoning for fake news detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2253–2262.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.
- 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.
- Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, et al. 2024. Llm-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 409–436.
- Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. [Fact-checking meets fauxtography: Verifying claims about images](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2099–2108, Hong Kong, China. Association for Computational Linguistics.

A Appendix

A.1 Dataset Example

Here is an example of dataset schema from MMCV:

Example

```
claim: Stoke City, a club that was
part of the top-tier league before
1992, was promoted to the highest
level of English football in 2018.
wiki_context: The Premier League
is the highest level of the English
football league system. Contested
by 20 clubs, it operates on a
system of promotion and relegation
with the English Football League
(EFL). Seasons usually run from
August to May, with each team
playing 38 matches: two against
each other, one home and one away.
Most games are played on weekend
afternoons, with occasional weekday
evening fixtures.
text_evidence: [
"f369ceelca92368c8b1ea564c5e41fc1"
]
image_evidence: []
table_evidence: [
"c120efadd518b5f32c11d40b456c8570"
]
label: SUPPORT
```

Additional examples of 1-hop, 2-hop, 3-hop, and 4-hop claims are listed in Table 6

A.2 Experiment Prompt

Claim Verification Prompt. To test MLLMs' claim verification performance under zero-shot settings, we follow (Geng et al., 2024b) and use the following prompt.

Prompt

Given a claim and evidence (which can be text, table, or an image), determine whether the claim is SUPPORT or REFUTE by the evidence.

Use the following format to provide your answer:

Prediction: [True or False]
Explanation: [put your evidence and step-by-step reasoning here]
Confidence Level: [please show the percentage]

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

Claim Generation Prompt. We use the following prompt to convert multimodal QA pairs into claim candidates:

Prompt

You are an expert in converting question-answers into claims. For example: Question: Telos was an album by a band who formed in what city? Answer: Indianapolis. Claim: Telos was an album by a band formed in Indianapolis.

Convert the question-answer into claim. Return only the claim and nothing else.

Claim Modification Prompt. We use the following prompt to modify the claim candidates:

Prompt

Generate a multi-hop specific claim based on the given general claim and Wikipedia context. The specific claim should:

Incorporate information from Wikipedia context.
Provided context should always be factually correct.
Obscure key information by:
a) Replacing one or two central entities with related fact using the Wikipedia context.
b) Alluding to critical details without explicitly stating them.
Claim should be short and concise.
For example:

- General Claim: The Mona Lisa is a famous painting by Leonardo da Vinci.
- Wikipedia Context: The Mona Lisa is a half-length portrait painting by Italian artist Leonardo da Vinci. Considered an archetypal masterpiece of the Italian Renaissance, it has been described as the best known, the most visited, the most written about, the most sung about, the most parodied work of art in the world. The painting's novel qualities include the subject's enigmatic expression, the monumentality of the composition, the subtle modelling of forms, and the atmospheric illusionism. It is housed in the Louvre Museum in Paris, where it was first put on display in 1797.
- Specific Claim: The Mona Lisa is a half-length portrait painting created by Italian artist who is considered as archetypal masterpiece of the Italian Renaissance.

Claim Refinement Prompt. We use the following prompt to refine the claim candidates:

Prompt

You are tasked with improving a claim focusing on three key areas: Fluency, Correctness, and Clearness. Your goal is to enhance the text while maintaining its original meaning and intent.

Improvement Criteria:

Fluency:

1. Review the text for grammar, syntax, and punctuation errors.
2. Rephrase any awkward or unnatural sentences to make the text flow more smoothly.
3. Ensure that the text reads naturally and is easy to follow.

Correctness:

1. Verify the factual accuracy of the content and correct any errors.
2. Ensure that the text adheres to the prompt's instructions.
3. Clarify any ambiguities and correct any inconsistencies in the information presented.

Clearness:

1. Simplify complex sentences or ideas to make the text easier to understand.
2. Improve the organization of ideas to enhance readability.
3. Ensure that the message is conveyed clearly and effectively, eliminating any confusion or ambiguity.

Final Output:

Once you have made the necessary improvements, provide the revised text. Ensure that the improved version is more fluent, accurate, and clear than the original while preserving the original meaning and intent.

Example Improvement:

Original Claim: "The results of the survey was very positive, with many respondents saying that they would recommend the service to others, however, some were also mentioned issues with the customer support."

Improved Claim: "The survey results were overwhelmingly positive, with many respondents stating they would recommend the service to others. However, some also noted issues with customer support."

A.3 Annotation Guidelines

We ask our annotators to score the quality of the claim from three aspects: fluency, correctness, and clearness. Here is the detailed guidelines provided to the human annotators.

Guidelines

▷ **Scoring Criteria:**

Fluency: Rate on a scale of 1-4.

Correctness: Rate on a scale of 1-3.

Clearness Rate on a scale of 1-3.

▷ **Fluency (1-4):**

4: Excellent - Reads naturally, no awkward phrasing.

3: Good - Mostly smooth, minor phrasing issues.

2: Fair - Several awkward phrases or constructions.

1: Poor - Difficult to read, very unnatural phrasing.

▷ **Correctness (1-3):**

3: Fully correct - All information is accurate.

2: Partially correct - Some information is accurate, some errors.

1: Incorrect - Significant factual errors or misrepresentations.

▷ **Clearness (1-3):**

3: Very clear - Easy to understand, no ambiguity.

2: Somewhat clear - Some parts may be confusing or ambiguous.

1: Unclear - Difficult to understand the intended meaning.

A.4 Crowd Worker Interface

We use SMART (Chew et al., 2019), an open-source project designed to help data scientists and research teams efficiently build labeled training datasets for supervised machine learning tasks. Figure 4 shows an example of the worker interface during scoring procedure.

#H

Claim

Evidence


1

Claim: Marisa Coughlan played the role of Chante Lefort on television in 1996.

	High Society	Dana	Episode: "Nip and Tuck"
1996	The Guilt	Kendall Cornell	Episode: "Dean's Office"
	The Burning Zone	Chante Lefort	Episode: "Blood Covenant"

2

Claim: The driver seen signing autographs outside had a significant points total during a specific race in 2001 while competing for a well-known team in stock car racing.



2001	NASCAR Winston Cup Series	Hendrick Motorsports	36	6	18	24	6	5112	1st
2002	NASCAR Winston Cup Series	Hendrick Motorsports	36	3	13	20	3	4607	4th

3

Claim: The Green Bay Packers were one of the two teams that played in the first Super Bowl and also faced the New York Giants at MetLife Stadium during the 2013 regular season.

Doc A: The first AFL-NFL World Championship Game in professional American football, known retroactively as Super Bowl I and

Doc B: The National Football League (NFL) champion Green Bay Packers defeated the American Football League (AFL) champion Kansas City Chiefs

Table: Not Included Here

4

Claim: The team that was promoted to the Premier League in the 2018-19 season received a higher accolade in the Third Division PFA Team of the Year during the 1980s than a club renowned for its West London rivalries.

Doc A: Manchester City are the defending champions. Wolverhampton Wanderers, Cardiff City and Fulham join as the promoted clubs from the 2017–18 EFL Championship.

Doc B: ... They will replace West Bromwich Albion, Swansea City and Stoke City who were relegated to the 2018–19 EFL Championship. ...

Table: Not Included Here

Image: Not Included Here

Table 6: Examples of 1-hop, 2-hop, 3-hop and 4-hop claims from MMCV.

Annotate Data
History
Fix Skew
Requires Adjudication 0
Discarded Data

Skip

Adjudicate

Text to Label

Wolverhampton Wanderers, a club that competed in the top tier, was relegated to the second division of the English football league system last year.

Respondent Data

Edit

Unnamed: 0

Evidence: The Premier League is the highest level of the English football league system. Contested by 20 clubs, it operates on a system of promotion and relegation with the English Football League (EFL). Seasons usually run from August to May, with each team playing 38 matches

FLU-4

FLU-3

FLU-2

FLU-1

Figure 4: UI for human annotators.





	1 Hop	2 Hop	3 Hop	4 Hop
GPT-4 o	<p>Claim: Durham, North Carolina, features a water tower located far from any building's chimney. True Label: REFUTE.</p> 	<p>Claim: Jakub Janda, a prominent figure in ski jumping, represents a country recognized by its flag that does not include a blue triangle. True Label: REFUTE</p> 	<p>Claim: The team that a player transferred from to Real Betis in the 2018–19 season, which features a flagpole in their logo, is not Real Sociedad. True Label: REFUTE</p> 	<p>Claim: The opponent in the regular season schedule of the 1991 Chicago Bears, known for their human logo, was a different team from which Roman Reigns had long careers in the NFL. True Label: REFUTE.</p> 
	<p>Predicted Label: SUPPORT Explanation: The image shows a water tower in Durham, North Carolina, which appears to be situated away from any visible building chimneys. The ... tower is located far from any building's chimney. Confidence Level: 85%</p>	<p>Predicted Label: SUPPORT Explanation: Claudio Pizarro was born in October 1978, which makes him 45 years old as of October 2023. If the unnamed Peruvian footballer is not ... footballer, the claim that Pizarro is older is supported. Confidence Level: 85%</p>	<p>Predicted Label: SUPPORT Explanation: China does indeed host numerous kite festivals throughout the year, including the Weifang International Kite Festival, which is a ... like Latin America, Southeast Asia, and Africa. Confidence Level: 95%</p>	<p>Predicted Label: SUPPORT Explanation: The image shows a CASA C-212 Aviocar aircraft, which is a turboprop aircraft. Turboprop aircraft are powered by propellers. Confidence Level: 95%</p>

Figure 5: Error Analysis: Gemini

	1 Hop	2 Hop	3 Hop	4 Hop
Gemini	<p>Claim: The background of the American legal drama series created by David E. Kelley is clear, especially concerning its exploration of legal ethics versus personal morality, which aimed to reinforce a more romanticized depiction of the legal system. True Label: REFUTE.</p> 	<p>Claim: Rochdale finished higher in the final league table of the 1994–95 season compared to the team known for its tree emblem, which has a history of competing in lower divisions. True Label: REFUTE</p> 	<p>Claim: In the film directed by a British filmmaker unknown for his collaboration with Walt Disney Studios, a Jamaican actor infamous for his role in a drama appears alongside two others in a car. True Label: REFUTE.</p> 	<p>Claim: The Argentine Naval Prefecture operates CASA C-212 Aviocar aircraft, which are equipped with propellers. True Label: REFUTE.</p> 
	<p>Predicted Label: SUPPORT Explanation: Extensive research on Scottish football databases and player profiles reveals no record of a player named Mark Connolly ... It's highly unlikely an unknown team would have achieved a significant victory over them. Confidence Level: 95%</p>	<p>Predicted Label: SUPPORT Explanation: Confidence Level: 100%</p>	<p>Predicted Label: SUPPORT Explanation: Chinese is widely spoken in Hong Kong: Cantonese, a variety of... is the primary language for 99.87% of the population in Northern Ireland, which is factually correct. Confidence Level: 99%</p>	<p>Predicted Label: SUPPORT Explanation: FA Cup Match: Wolverhampton Wanderers did indeed play ... finished second in the Championship, the second tier of English football, securing automatic promotion. Confidence Level: 95%</p>

Figure 6: Error Analysis: GPT4-o



LLaVA	1 Hop	2 Hop	3 Hop	4 Hop
	<p>Claim: The Lake Superior State Lakers men's ice hockey team features a star in its logo. True Label: REFUTE</p> 	<p>Claim: The poster for the new Wizard of Oz movie features five characters. True Label: REFUTE</p> 	<p>Claim: Aaron Gorrell is the player who transferred to the Catalans Dragons during the 2008 season, having joined in October 2007, and is depicted holding a trophy. True Label: REFUTE</p> 	<p>Claim: The Thüringenliga is one of several regions listed under "Promotion To" in the Tier VII-Landesligas that partially extends beyond the country's borders and was designated under "Promotion to" when the "Relegation to" listing was not Kreisoberliga. True Label: REFUTE</p> 
	<p>Predicted Label: SUPPORT Explanation: The Lake Superior State Lakers ... The team's logo features a star, which is a common symbol in college sports. ... Therefore, it is reasonable to assume that the team's logo features a star. Confidence Level: 90%</p>	<p>Predicted Label: SUPPORT Explanation: The poster for the new Wizard of Oz movie features five characters, which is evident from the image. ... The poster's design and the choice of characters indicate that the movie is likely to be a faithful ... Confidence Level: 90%</p>	<p>Predicted Label: SUPPORT Explanation: Aaron Gorrell is a rugby league player who ... There is no evidence to refute this claim, and it is supported by the fact that Gorrell was a player for the Catalans Dragons during the 2008 season. Confidence Level: 90%</p>	<p>Predicted Label: SUPPORT Explanation: The Thüringenliga is indeed one of the regions listed under "Promotion To" in the Tier VII-Landesligas that partially ... that it is a regional league that can be promoted to higher levels of the German football league system. Confidence Level: 90%</p>

Figure 7: Error Analysis: LLaVA