RAGAR, Your Falsehood Radar: RAG-Augmented Reasoning for Political Fact-Checking using Multimodal Large Language Models

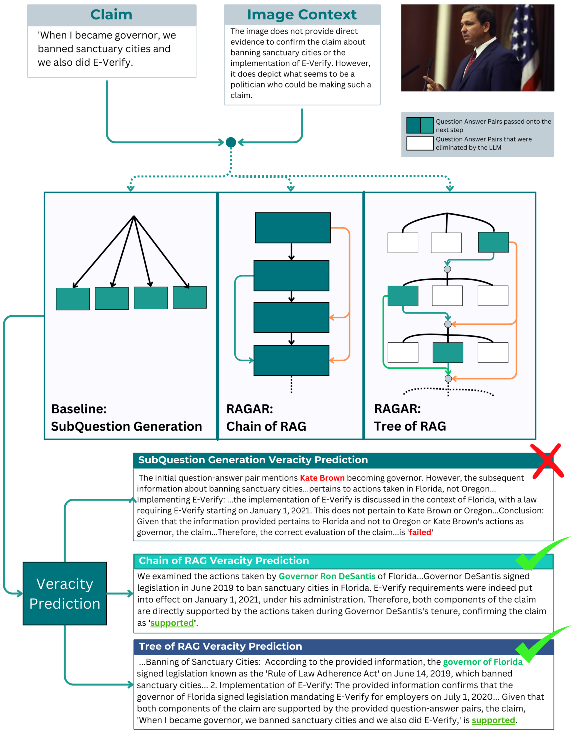
Mohammed Abdul Khaliq[1,3], **Paul Yu-Chun Chang**[2]*,
Mingyang Ma[2], **Bernhard Pflugfelder**[2], **Filip Miletić**[1]*
[1]Institute for Natural Language Processing, University of Stuttgart,
[2]appliedAI Initiative GmbH, [3]appliedAI Institute for Europe gGmbH
`{mohammed.abdul-khaliq, filip.miletic}@ims.uni-stuttgart.de`,
`{p.chang, m.ma, b.pflugfelder}@appliedai.de`

Abstract

The escalating challenge of misinformation, particularly in political discourse, requires advanced fact-checking solutions; this is even clearer in the more complex scenario of multimodal claims. We tackle this issue using a multimodal large language model in conjunction with retrieval-augmented generation (RAG), and introduce two novel reasoning techniques: Chain of RAG (CoRAG) and Tree of RAG (ToRAG). They fact-check multimodal claims by extracting both textual and image content, retrieving external information, and reasoning subsequent questions to be answered based on prior evidence. We achieve a weighted F1-score of 0.85, surpassing a baseline reasoning technique by 0.14 points. Human evaluation confirms that the vast majority of our generated fact-check explanations contain all information from gold standard data.

1 Introduction

In the age of digital information, rapid dissemination of news, both genuine and fabricated, has become a defining feature of public discourse. The phenomenon of fake news – which more precisely denotes misinformation, disinformation, or a combination of both (Aïmeur et al., 2023) – is particularly prevalent on social media: false information spreads six times faster than the truth on platforms like Twitter (Vosoughi et al., 2018). This trend poses a critical challenge to the democratic process since it makes voters increasingly prone to making decisions based on incorrect information. The matter is further aggravated by visual information, which provides yet another widespread and consequential source of fake news. For instance, fake news stories that include images spread further than those containing only text (Zannettou et al., 2018).


Claim
Image Context
Question Answer Pairs passed onto the next step
Question Answer Pairs that were eliminated by the LLM
Baseline:
SubQuestion Generation
RAGAR:
Chain of RAG
RAGAR:
Tree of RAG
Veracity Prediction
SubQuestion Generation Veracity Prediction
Chain of RAG Veracity Prediction
Tree of RAG Veracity Prediction

'When I became governor, we banned sanctuary cities and we also did E-Verify.

The image does not provide direct evidence to confirm the claim about banning sanctuary cities or the implementation of E-Verify. However, it does depict what seems to be a politician who could be making such a claim.

The initial question-answer pair mentions **Kate Brown** becoming governor. However, the subsequent information about banning sanctuary cities...pertains to actions taken in Florida, not Oregon... Implementing E-Verify: ...the implementation of E-Verify is discussed in the context of Florida, with a law requiring E-Verify starting on January 1, 2021. This does not pertain to Kate Brown or Oregon...Conclusion: Given that the information provided pertains to Florida and not to Oregon or Kate Brown's actions as governor, the claim...Therefore, the correct evaluation of the claim...is '**failed**'

We examined the actions taken by **Governor Ron DeSantis** of Florida...Governor DeSantis signed legislation in June 2019 to ban sanctuary cities in Florida. E-Verify requirements were indeed put into effect on January 1, 2021, under his administration. Therefore, both components of the claim are directly supported by the actions taken during Governor DeSantis's tenure, confirming the claim as '**supported**'.

...Banning of Sanctuary Cities: According to the provided information, the **governor of Florida** signed legislation known as the 'Rule of Law Adherence Act' on June 14, 2019, which banned sanctuary cities... 2. Implementation of E-Verify: The provided information confirms that the governor of Florida signed legislation mandating E-Verify for employers on July 1, 2020... Given that both components of the claim are supported by the provided question-answer pairs, the claim, 'When I became governor, we banned sanctuary cities and we also did E-Verify,' is **supported**.

Figure 1: An overview of the fact-checking pipeline contrasting the baseline Sub-Question Generation approach from the Chain of RAG and Tree of RAG approach followed by veracity prediction and explanation.

A potential solution to these issues is provided by automated fact-checking systems. They have benefited from the development of large language models (LLMs), leading to improvements in detection, labeling, and generation of veracity explanations (Das et al., 2023). More recently, multimodal approaches have complemented textual information with image representations to assess their cross-modal consistency and unified embedding representations (Yao et al., 2023a). Another active line of research deploys retrieval-augmented generation (RAG), whereby LLMs access up-to-date external information at inference time. They convert the input claim into phrase queries, pass them onto a search engine, and use the retrieved information to assess veracity (Asai et al., 2024; Zeng and Gao,

---

*Corresponding authors.

Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER), pages 280–296
November 15, 2024 ©2024 Association for Computational Linguistics

2024). It however remains to be determined if more elaborate reasoning techniques can be beneficial in this setting. Moreover, RAG-based approaches have so far mostly been applied to text. This raises the additional question of their use in the more challenging scenario of multimodal fact-checking.

Addressing this gap, we introduce RAGAR – RAG-Augmented Reasoning techniques, which we apply to multimodal fact-checking in the political domain (see Figure 1 for a high-level overview). We rely on a multimodal LLM to verbalize the textual and visual elements of a claim, and use RAG responses to motivate successive steps in determining veracity. The system is underpinned by elaborate reasoning strategies instantiated in two distinct approaches: Chain of RAG (CoRAG) and Tree of RAG (ToRAG). We evaluate them using a multimodal fact-checking dataset as well as human annotation of generated explanations.

Our contributions are as follows. (1) We introduce two novel reasoning techniques for multimodal fact-checking, reaching a weighted F1-score of 0.85. (2) We provide two complementary strategies for multimodal input by verbalizing image content during claim generation and using image captions as evidence during retrieval. (3) We conduct a multi-rater annotation of fact-check explanations, showing that the vast majority of them include all information from the gold standard. To our knowledge, this is the first study to incorporate multimodal LLMs in a RAG-based reasoning approach applied to multimodal fact-checking for the political domain.

2 Related Work

2.1 Retrieval-Augmented Generation (RAG) for Fact-Checking

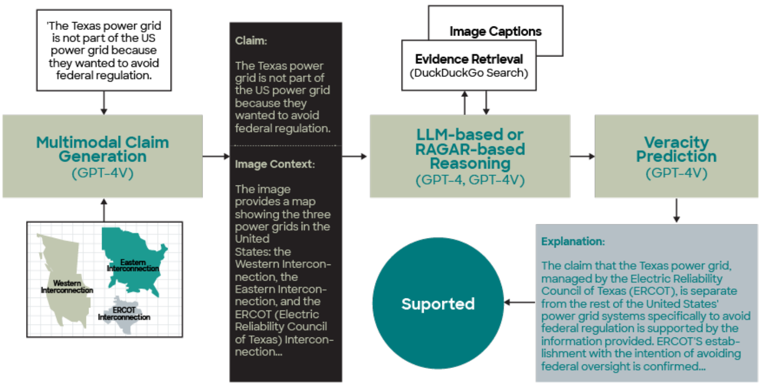
To combat hallucination in text generation, current fact-checking pipelines often implement a RAG approach, wherein an LLM retrieves data from external sources to enhance its response and move past its knowledge cutoff. Peng et al. (2023) present LLM-Augmenter, which combines external knowledge integration and automated feedback mechanisms. Chern et al. (2023) assess the factuality of LLM-generated text on multiple tasks and domains, e.g. for Knowledge Based Question Answering they use Google Search API to extract relevant knowledge and then parse the result. Pan et al. (2023) rely on LLM's in-context learning, and use Chain of Thought (Wei et al., 2022) reasoning to guide the model in complex tasks such as fact-checking on the web. Zhang and Gao (2023) propose Hierarchical Step-by-Step (HiSS) prompting, which splits a claim into sub-claims, creating a hierarchy, and verifies each one through multiple question-answering steps using web-retrieved evidence. Xu et al. (2023) propose SearChain. It creates a Chain of Query (CoQ) reasoning chain, where each question follows from the knowledge gathered in the previous question; uses information retrieval (IR) to verify the answer at each node; and prompts the LLM to indicate missing information, which is handled by an IR call.

Our RAGAR approaches are conceptually similar, but they use a more sophisticated reasoning framework with multiple rounds of sequential question-answering, elimination (in the case of ToRAG), and verification. We also extend domain coverage through multimodality, and propose a zero-shot (rather than few-shot) approach.

2.2 Multimodal Fact-Checking using LLMs

Multimodality is generally underexplored in fact-checking (Alam et al., 2022), but several recent approaches have been proposed. Guo et al. (2023) use LLM-agnostic models to generate textual prompts from images and then guide LLMs in generating responses to Visual Question Answering queries. Yao et al. (2023a) construct a multimodal dataset using fact-checking websites, and then develop a fact-checking and explanation generation pipeline. It encodes and reranks each sentence in the document corpus in relation to the claim, and uses a CLIP (Radford et al., 2021) encoding for images; the similarity between an input claim and the provided images is then computed. An attention model is used for multimodal claim verification, and BART (Lewis et al., 2020) for explanation generation. In concurrent research, Pan et al. (2024) propose the Chain of Action prompt. It splits an input query into sub-questions and uses a "Missing Flag" indicator to fill in or correct the answers generated by internal LLM knowledge via RAG.

Our RAGAR approaches similarly use a multimodal LLM (GPT-4V; OpenAI, 2023) to add context to the textual claim, but employ a different set of reasoning techniques. We furthermore introduce a multimodal RAG component during evidence retrieval, using captions of matching images to provide the LLM with relevant meta information.



'The Texas power grid is not part of the US power grid because they wanted to avoid federal regulation.

Claim:

The Texas power grid is not part of the US power grid because they wanted to avoid federal regulation.

Image Context:

The image provides a map showing the three power grids in the United States: the Western Interconnection, the Eastern Interconnection, and the ERCOT (Electric Reliability Council of Texas) Interconnection...

Explanation:

The claim that the Texas power grid, managed by the Electric Reliability Council of Texas (ERCOT), is separate from the rest of the United States' power grid systems specifically to avoid federal regulation is supported by the information provided. ERCOT'S establishment with the intention of avoiding federal oversight is confirmed...

Figure 2: A detailed overview of the Multimodal Fact-checking pipeline

3 Dataset

The aim of our study is to explore the potential of multimodal LLM-based RAG and reasoning for political fact-checking. Given the substantial computational and financial costs of running multimodal LLMs through multiple rounds of reasoning, we evaluate our approach on a well-controlled and balanced dataset, so as to minimize noise while maintaining the validity of our experiments.

We specifically rely on a carefully selected subset of the MOCHEG dataset (Yao et al., 2023a). MOCHEG provides 21,184 multimodal claims sourced from two fact-checking websites, PolitiFact[1] and Snopes.[2] Each instance contains an input claim extracted from the title of the fact-checking source, and an associated image extracted from the web page that addresses the claim. The dataset further provides a summary of the fact-check in the form of a "Ruling Outline", which we consider for evaluating LLM-generated explanations.

We start from the test set containing 2,007 multimodal claims and filter it in two steps. First, we select the 794 claims that were fact-checked by PolitiFact, since our focus is on political claims; by contrast, Snopes provides fact-checks for a variety of domains. Second, we filter this set down to 300 test samples randomly selected from the *supported* and *refuted* classes, for a balanced final dataset with 150 multimodal claims in each of the two classes.

In this process, we purposefully discard the *NEI* (Not Enough Information) instances. During the creation of MOCHEG, some ambiguous cases were outright discarded, while the labels *mixture*, *unproven*, and *undetermined* were aggregated under *NEI*. This class is potentially unstable in two respects: fact-checking websites update their labels as new evidence emerges (Yao et al., 2023a), which by definition affects this class more prominently; and the fact-checking intentions behind mixed labels such as *half-true* and *mixture* are comparatively unclear, leading prior studies to exclude them (e.g. Vo and Lee, 2019). We adopt the same decision given our focus on an initial validation of novel reasoning techniques.

Although we only retain instances that are unambiguous in the dataset, our model may still struggle to retrieve information of sufficient quality to fact-check them. We account for this by allowing it to generate a *failed* label when it fails to retrieve relevant information. We reserve an extension of our study to the *NEI* class, as well as the connected issue of improving retrieval quality, for future work.

4 Multimodal Fact-Checking Pipeline

Our fact-checking pipeline comprises four parts: (i) Multimodal Claim Generation, which analyzes both the textual claim and associated image to formulate a new claim incorporating both; (ii) Multimodal Evidence Retrieval, which extracts evidence from the web for a question posed by the LLM; (iii) LLM-based and RAG-augmented Reasoning for fact-checking, our reasoning approach to fact-check a claim; and (iv) Veracity Prediction and Explanation. The pipeline is shown in Figure 2.

4.1 Multimodal Claim Generation

Given an input claim as text, an associated image, and the date of the claim, the claim generation module generates a response verbalizing the information contained in both the textual claim and the image. We use GPT-4V as our multimodal LLM given its strong performance across tasks. Note that our aim is not to determine the best-performing

---

[1] `https://www.politifact.com/`

[2] `https://www.snopes.com/`

model on our task, but rather to evaluate different reasoning techniques. We therefore use the same model across experiments.

The generated response is divided into two sections: *claim*, which contains the original text claim; and *image context*, which contains the details relevant to the claim extracted from the image by GPT-4V. The *image context* expands on the information from the textual claim by e.g. identifying the speaker that the claim is quoting, extracting numerical information from figures, and highlighting relevant textual data mentioned in the image. More generally, the contextualization provides details on whether the image is relevant to the text claim.

While directly encoding images is a potential alternative to our approach, we decide against it to allow our Chain of RAG and Tree of RAG approach to be multimodal-agnostic. This decision ensures that our reasoning methods can also be replicated with LLMs that are not inherently multimodal. Multimodal Claim Generation is the only section of our pipeline requiring a multimodal LLM; all remaining parts, including our RAGAR approaches, can be implemented using other LLMs and possibly extended to different tasks.

4.2 Multimodal Evidence Retrieval

The fact-checking questions generated by the LLM-based or RAG-augmented reasoning techniques serve as input for the multimodal evidence retrieval module. It helps answer each question by retrieving relevant text snippets from websites and further analyzing details associated with the image.

The query to the multimodal evidence retrieval is a question generated by an LLM-based or RAGAR-based reasoning technique (presented in detail in Section 4.3). For text-based evidence retrieval, we use the DuckDuckGo Search tool provided by LangChain[3]. We retrieve the top 10 results from the API and use them to answer the question. We temporally restrict the search by only collecting articles published in the two years before the claim was fact-checked by PolitiFact, so as to provide the LLM with facts relevant to the time-frame of the fact-check. To mimic a real-time fact-checking scenario, we remove search results that originate from `www.politifact.com`, `www.snopes.com`, and `www.factcheck.org`, since it is likely that they already contain answers to the claim and would thus impact the fairness of the experiment. We also remove the following social media websites due to potentially biased or unreliable information: `www.facebook.com`, `www.tiktok.com`, `www.twitter.com` and `www.youtube.com`.

Most images in our dataset contain faces of politicians, pictures from political events, government buildings etc. In such cases, the image itself may not provide much additional information beyond the text claim. However, it is useful to determine the metadata associated with the image, which may indicate when or where the claim was made. For this purpose, we use SerpAPI[4] to conduct a reverse image search over the images associated with the claims. We extract the captions for the images from the first 10 results and use them as additional information for GPT-4V. This allows the model to not only analyze the image when answering an image-based question, but also incorporate meta-information about it and in that way better contextualize the answer. We demonstrate a few examples of this in Appendix A.3.

4.3 LLM-Based and RAG-Augmented Reasoning for Fact-Checking

4.3.1 Baseline: Sub-questions with Chain of Thought at Veracity Prediction (SubQ+CoT$_{\text{VP}}$)

As a baseline reasoning-based approach, we employ sub-question generation followed by Chain of Thought veracity prediction (SubQ+CoT$_{\text{VP}}$). This baseline is based on recent approaches to fact-checking relying on LLMs (Pan et al., 2023; Chern et al., 2023) as discussed in Section 2.1. We adapt the approach to handle multimodal claims as well.

4.3.2 RAG-Augmented Reasoning: Chain of RAG (CoRAG)

The first novel reasoning approach we propose is Chain of RAG (CoRAG). It builds upon general RAG approaches by using sequential follow-up questions – augmented from the RAG response – to retrieve further evidence. In other words, we follow a decomposed setup, guiding the LLM towards asking questions based on the previously generated question-answer pairs. The “Chain” in “Chain of RAG” is thus to be interpreted as a chain of question-answer pairs that are iteratively generated. This is unlike the traditional Chain of Thought, wherein a single prompt handles the entire process of creating questions, answers, and follow-up ques-

---

[3] `https://www.langchain.com/`

[4] `https://serpapi.com/`

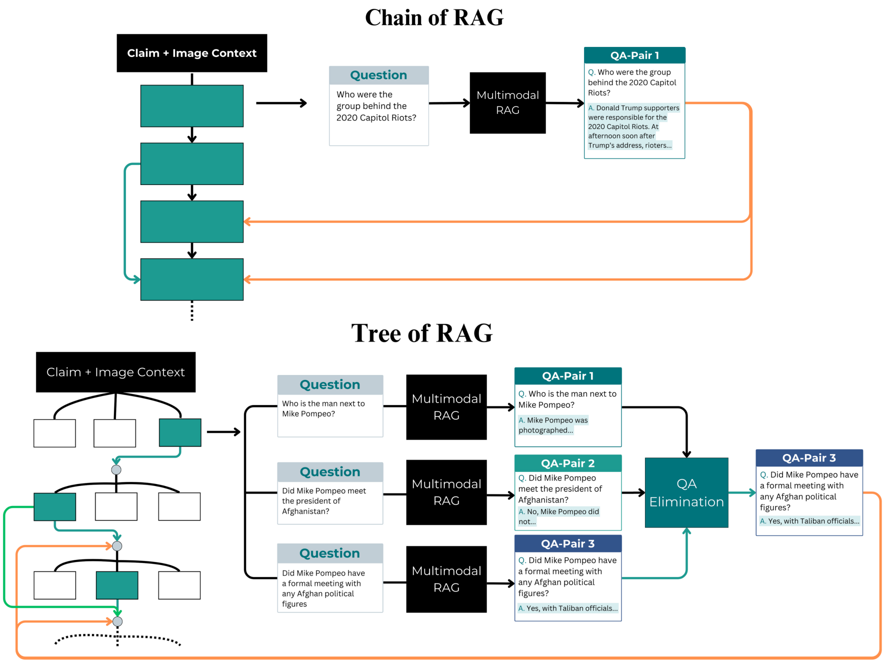

Chain of RAG
Claim + Image Context
Question
Who were the group behind the 2020 Capitol Riots?
Multimodal RAG
QA-Pair 1
Q. Who were the group behind the 2020 Capitol Riots?
A. Donald Trump supporters were responsible for the 2020 Capitol Riots. At afternoon soon after Trump's address, rioters...
Tree of RAG
Claim + Image Context
Question
Who is the man next to Mike Pompeo?
Question
Did Mike Pompeo meet the president of Afghanistan?
Question
Did Mike Pompeo have a formal meeting with any Afghan political figures
Multimodal RAG
Multimodal RAG
Multimodal RAG
QA-Pair 1
Q. Who is the man next to Mike Pompeo?
A. Mike Pompeo was photographed...
QA-Pair 2
Q. Did Mike Pompeo meet the president of Afghanistan?
A. No, Mike Pompeo did not...
QA-Pair 3
Q. Did Mike Pompeo have a formal meeting with any Afghan political figures?
A. Yes, with Taliban officials...
QA Elimination
QA-Pair 3
Q. Did Mike Pompeo have a formal meeting with any Afghan political figures?
A. Yes, with Taliban officials...

Figure 3: Chain of RAG and Tree of RAG pipeline

Algorithm 1 Chain of RAG (CoRAG)

```
Input: Claim $C$, Image Context $I$, Image Captions $IC$
$Q \leftarrow$ GenerateFirstQuestion($C, I$)
QAPairs ← []                                          ▷ Initialize an empty list for Q-A pairs
$counter \leftarrow 0$
$followUpNeeded \leftarrow$ True
while $counter < no\_of\_steps$ and $followUpNeeded$ do
    if QuestionAboutImage($Q$) then
        $A \leftarrow$ ImageQA($Q, I, IC$)            ▷ Using image, question, and captions
    else
        $A \leftarrow$ WebQA($Q$)                     ▷ Standard evidence retrieval
    end if
    QAPairs.append(($Q, A$))                          ▷ Store the Q-A pair of this iteration
    $followUpNeeded \leftarrow$ FollowupCheck($Q, A$)
    if $followUpNeeded$ then
        $Q \leftarrow$ FollowupQuestion($QAPairs$)
    end if
    $counter \leftarrow counter + 1$
end while
return QAPairs                                        ▷ Returns the list of Q-A pairs
```

tion in one go. Moreover, CoRAG follows a zero-shot approach, i.e. the LLM is not provided with any example question-answer pairs to influence the reasoning process. An overview of the process is provided in Algorithm 1 as well as Figure 3.

The input to the CoRAG module is the *claim* and *image context* from the multimodal claim generation module (§4.1). The LLM is first prompted to generate a question that is intended to answer an aspect of the claim. The generated question is passed to the multimodal evidence retriever (§4.2), which obtains evidence to inform the RAG answer. Once the answer is generated, the CoRAG process undergoes a follow-up check (effectively an early termination check). The follow-up check prompt (see Appendix A.5) takes as input the LLM-generated claim as well as all the generated question-answer pair(s), and checks whether enough information has been gathered to answer the claim. If the response from the follow-up check is "True", it asks a follow-up question. The follow-up question is intended to ask for further information, building on top of the previous question-answer pairs such that the claim can be fully addressed.

A follow-up check occurs after each question-answer generation step. If the follow-up check prompt finds sufficient evidence in the questions and answers generated up until that point, it terminates and passes the evidence to the veracity prediction and explanation generation module. We also set a constraint of a maximum of six questions, after which the CoRAG process terminates even if it does not have enough evidence for the fact-check. We determined this threshold in preliminary experiments on 80 samples, which indicated that this was the highest number of question-answering steps required for the LLM to obtain enough information to address even the more challenging claims.

4.3.3 RAG-Augmented Reasoning: Tree of RAG (ToRAG)

In a similar way to how a traditional Tree of Thought (Yao et al., 2023b) extends Chain of Thought through branching, Tree of RAG (ToRAG) extends our CoRAG approach by creating question branches at each reasoning step. The best question-answer branch is selected at each step. An overview is provided in Algorithm 2 as well as Figure 3.

The input to the ToRAG module is the *claim* and *image context* from the multimodal claim generation module (§4.1). Upon receiving this input, the ToRAG approach branches into three, each branch asking a unique question to fact-check the claim.

Once the three starting questions have been generated, the ToRAG approach uses the evidence retriever (§4.2) to obtain information and generate answers for each question. The three question-answer pairs are then passed into an elimination prompt, from which only one question-answer pair is chosen as candidate evidence. The model is prompted to perform this elimination based on relevance, detail, additional information, and answer confidence (see Appendix A.6).

The candidate evidence then serves as the basis for the follow-up question. Three follow-up questions are generated simultaneously based on the candidate evidence. The evidence retriever fetches answers to these questions, and the LLM generates the answers. New candidate evidence is chosen by the elimination prompt and is added to the existing list of candidate evidence. This list, therefore, stores only the best of the three question-answer pairs obtained at each step. Upon gathering sufficient information to fact-check the claim as determined by the follow-up check prompt or reaching a maximum of six candidate evidence question-answer pairs, the ToRAG process terminates, and the list of candidate evidence is passed to the veracity prediction and explanation generation module. A few examples of the question-answer pairs generated by our LLM-based and RAG-augmented reasoning approaches can be seen in Appendix A.4.

4.4 Veracity Prediction and Explanation

The veracity prediction and explanation module (henceforth referred to as "veracity prediction" for brevity) generates a veracity label of *supported* or *refuted* based on the information available in the question-answer pairs. Moreover, it generates a *failed* label when it deems to have insufficient information in the question-answer pair to either support or refute the claim.

We experiment with three variants of veracity prediction prompts (see Appendix A.7). (i) The standard veracity prompt ($\text{Standard}_{\text{VP}}$) takes the claim and evidence pairs as input, and outputs the veracity rating and the explanation without any induced reasoning. (ii) The zero-shot Chain of Thought veracity prediction prompt (CoT_{VP}) uses the "Let's think step by step" phrase to guide the model to follow a chain of thought reasoning approach. (iii) The Chain of Verification (Dhuliawala et al., 2023) veracity prediction prompt (CoVe) first constructs verification questions based on the LLM-generated fact-checked explanation. The answers to these questions are generated using RAG, and are passed – along with the LLM-generated fact-check – to a correction check prompt. In case of corrections to the original LLM-generated fact-check, a new fact-check is generated along with a new veracity label if necessary. The CoVe veracity prediction approach is thus able to verify the fact-checked explanation generated by the CoRAG and ToRAG methods with the intended goal of capturing and correcting hallucination.

5 Evaluation and Results

We now present two evaluations employed across the set of 300 multimodal claims. In Section 5.1, we analyze system performance based on the correctness of veracity predictions. In Section 5.2, we zoom into explanation generation by conducting a human annotation study to compare the generated and gold explanations.

5.1 Correctness of Veracity Predictions

In this evaluation setup, we categorize the predictions into two primary outcomes: correct or incorrect. Specifically, when the language model's prediction matches the actual label (for instance, predicting *supported* when the actual rating is *supported*), the prediction is deemed correct. Conversely, if the model predicts *refuted* or *failed* when the actual rating is *supported*, the prediction is con-

Algorithm 2 Tree of RAG (ToRAG)

```
Input: Claim $C$, Image Context $I$, Image Captions $IC$
BestQAPairs ← []                                    ▷ Initialize an empty list for best Q-A pairs
Questions ← GenerateFirstQuestions($(C, I)$)        ▷ Generates three questions
$counter \leftarrow 0$
$followUpNeeded \leftarrow$ True
while $counter < no\_of\_steps$ and $followUpNeeded$ do
    QAPairs ← []                                    ▷ Initializes an empty list for question-answer pairs
    for $Q$ in Questions do
        if QuestionAboutImage($(Q)$) then
            $A \leftarrow$ ImageQA($(Q, I, IC)$)       ▷ Using image, question, and captions
        else
            $A \leftarrow$ WebQA($(Q)$)               ▷ Standard evidence retrieval
        end if
        QAPairs.append($((Q, A))$)
    end for
    (BestQ, BestA) ← QAElimination(QAPairs)
    BestQAPairs.append((BestQ, BestA))              ▷ Stores the best Q-A pair of this iteration
    $followUpNeeded \leftarrow$ FollowupCheck(BestQAPairs)
    if $followUpNeeded$ then
        Questions ← GenerateFollowupQuestions(BestQAPairs)   ▷ Generates three follow-up questions
    else
        break
    end if
    $counter \leftarrow counter + 1$
end while
return BestQAPairs                                  ▷ Returns all collected best Q-A pairs
```

APPROACHES	SUPPORTED (F1)	REFUTED (F1)	# FAILED	WEIGHTED F1
SubQ + CoT$_{\text{VP}}$	0.66	0.77	50 | 22	0.71
CoRAG + Standard$_{\text{VP}}$	0.74	0.81	31 | 15	0.77
CoRAG + CoT$_{\text{VP}}$	0.73	0.82	38 | 14	0.77
CoRAG + CoT$_{\text{VP}}$ + CoVe	0.78	0.83	21 | 8	0.81
ToRAG + Standard$_{\text{VP}}$	0.82	**0.86**	16 | 5	0.84
ToRAG + CoT$_{\text{VP}}$	0.82	0.85	19 | 9	0.83
ToRAG + CoT$_{\text{VP}}$ + CoVe	**0.84**	**0.86**	9 | 4	**0.85**

Table 1: F1 Results of the Correctness of Veracity Predictions evaluation. The # FAILED column contains the number of *supported* | *refuted* claims that were predicted as *failed*.

sidered as incorrect. Table 1 shows the results of all of our approaches for this evaluation criterion.

The worst-performing approach is the SubQ+CoT$_{\text{VP}}$ baseline, with a weighted F1 of 0.71. The best-performing approach is ToRAG+CoT$_{\text{VP}}$+CoVe, with a weighted F1 of 0.85. The middle spot is occupied by the CoRAG implementations; the strongest among those is CoRAG+CoT$_{\text{VP}}$+CoVe, with a weighted F1 of 0.81. Regarding class-level performance, the scores are consistently higher for the *refuted* rather than *supported* class.

The SubQ+CoT$_{\text{VP}}$ baseline lags behind our RAGAR approaches by up to 0.14 weighted F1 points. We attribute its poor performance to the inability of the veracity prediction module (CoT$_{\text{VP}}$) to gain sequential and contextual information. Since the subquestions generated by SubQ+CoT$_{\text{VP}}$ are based solely on the claim, the answers queried during evidence retrieval do not follow from one another.

Amongst our RAGAR approaches, applying CoT$_{\text{VP}}$ to the question-answer pairs generated by either CoRAG or ToRAG approaches did not show improvement over Standard$_{\text{VP}}$. We attribute this to the very strong internal reasoning capabilities of GPT-4. However, we are able to improve performance by combining the CoVe approach, especially in the case of CoRAG. Incorporating CoVe with the result from CoRAG+CoT$_{\text{VP}}$ shows a performance improvement of 0.04 F1 points and especially improves the classification of *supported* claims. Incorporating CoVe on top of the ToRAG+CoT$_{\text{VP}}$ leads to an improvement, but overall minor and also less pronounced than for CoRAG. This indicates that the QA elimination prompt in ToRAG successfully eliminates erroneous or irrelevant question-answer pairs.

5.2 Evaluating Explanation Generation

We evaluate explanation generation by comparing the LLM-generated fact-checked explanation with the corresponding "Ruling Outline" from the MOCHEG dataset. We recruit three volunteer annotators, aged 21–24 and with near-native English proficiency. They are asked to rate the explanations generated by each of the approaches on a scale from

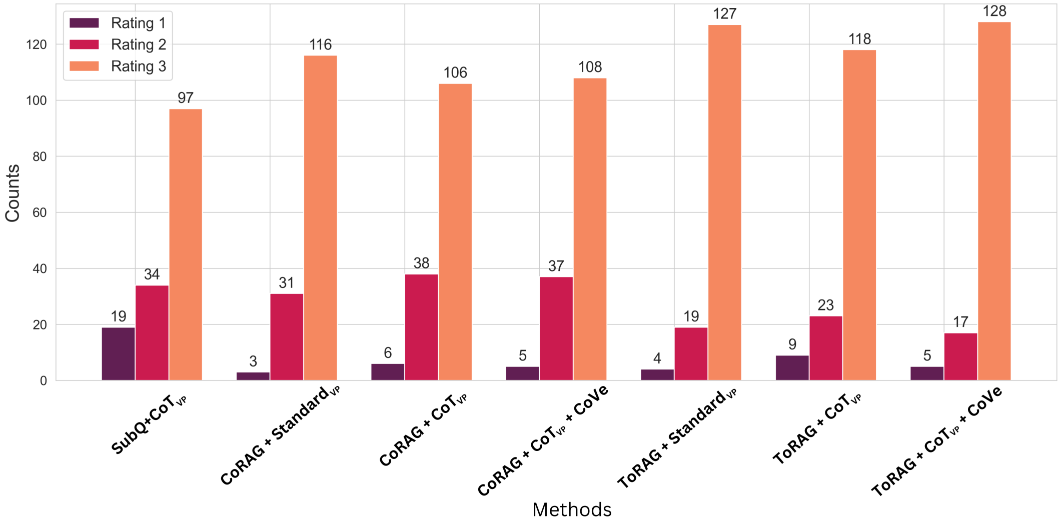


Figure 4: Number of 1/2/3 ratings received for explanations by each approach

1 to 3, where 3 indicates that all information in the gold explanation is present in the generated explanation, while 1 indicates that all information in the gold explanation is missing from the generated explanation. The complete annotation instructions are provided in Appendix A.1.

We randomly sample a set of 50 claims, divided into 25 supported and 25 refuted. For all annotated claims, the gold veracity label and the predicted veracity label match. We measure inter-annotator agreement using Krippendorff's α (Hayes and Krippendorff, 2007). The scores are in the range of 0.53 to 0.75 depending on the evaluated approach, with the mean at 0.60. We consider this to be sufficient agreement given the nature of the task.

As can be seen in Figure 4, the annotators provide a rating of 3 for an overwhelming majority of explanations generated across methods. This shows that the generated explanations indeed cover all the points noted in the PolitiFact fact-check. Additionally, the explanations generated by SubQ+CoT$_{\text{VP}}$ led to significantly more ratings of 1 than any other method, which indicates that it omitted or did not accurately elaborate on certain points.

Regarding class-level trends, explanations in the *supported* class are rated as 2 more often than those in the *refuted* class (see Appendix A.2). This indicates that certain information was missing from the generated explanation; more generally, this trend reflects the lower F1 scores on this class (§5.1), suggesting its higher difficulty. From a qualitative perspective, the annotators anecdotally reported that the generated explanations included some points from the PolitiFact ruling outline, but also provided additional information. Overall, however, the majority of the ratings being annotated as 3 across the different approaches lends credence to the quality of the explanation and to the efficacy of the underlying system in retrieving relevant evidence to fact-check the claim.

6 Conclusion

This paper introduces and tests two new methods for political fact-checking using large language models (LLMs): Chain of RAG (CoRAG) and Tree of RAG (ToRAG). These methods tackle misinformation in political discussions, focusing on multimodal claims, and show notable improvements over traditional fact-checking approaches that use sub-question generation with LLMs. CoRAG uses a step-by-step questioning strategy for thorough claim examination, while ToRAG extends upon this by following a branching strategy with evidence elimination thereby enhancing veracity prediction. We evaluate these methods in two ways. In terms of correctness of generated veracity label, we see an increase of 0.06-0.14 F1 points when using the RAGAR framework with Standard, CoT$_{\text{VP}}$, and CoVe veracity prediction prompts compared to the baseline SubQ+CoT$_{\text{VP}}$. For explanation generation, the quality of RAGAR-generated explanations was consistently rated higher than the baseline method. Our study shows that RAG-augmented reasoning (RAGAR) techniques are effective in multimodal political fact-checking, improving both the accuracy of veracity predictions and the quality of detailed fact-check explanations.

7 Limitations

We experimented with three tools for extracting relevant web results for natural language questions; DuckDuckGo Search, You.com[5] and Tavily AI[6]. Across the three tools, we notice that the search results may occasionally vary when prompted with the same questions multiple times. This variance in results, even though the question remains the same or similar, is problematic since it affects the final result and makes it hard to compare approaches. Additionally, due to budget constraints, we are unable to provide variance estimates requiring multiple runs of our RAGAR approaches. While we acknowledge the use of a closed-source LLM as a potential shortcoming due to comparatively more limited control over model behavior, we opted for the best-performing model available to us given the complexity of the addressed task. Finally, as also noted in the paper, our main aim was to assess the viability of novel reasoning techniques rather than retrieval quality, which led us to exclude *NEI* instances from our experimental setup. Further work extended to these cases is needed to more comprehensively understand the performance of our proposed approach.

8 Ethics Statement

We conducted an experimental study aimed at examining multimodal fact-checking by prompting LLMs, and note that some of the core steps of this approach may also be replicated by the general public. Our RAGAR approach obtained clear improvements over the examined baseline in the evaluation setup we defined. However, the experiments presented here are not sufficient to make general claims about the performance of our approach in other settings. Given the sensitive nature of political news in particular, we caution against using the RAGAR approach for general political fact-checking or implementing it on a large scale at this stage.

9 Acknowledgements

We thank the anonymous reviewers for their constructive feedback. Filip Miletić was supported by DFG Research Grant SCHU 2580/5-1.

References

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. FacTool: Factuality detection in generative AI – a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered NLP technology for fact-checking. *Information Processing & Management*, 60(2):103219.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Zixian Guo, Bowen Dong, Zhilong Ji, Jinfeng Bai, Yiwen Guo, and Wangmeng Zuo. 2023. Texts as images in prompt tuning for multi-label image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2808–2817.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1:77 – 89.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4v: A multimodal transformer for vision and language.

---

[5] `https://you.com/`

[6] `https://tavily.com/`

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Zhenyu Pan, Haozheng Luo, Manling Li, and Han Liu. 2024. Chain-of-action: Faithful and multimodal question answering through large language models. *arXiv preprint arXiv:2403.17359*.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-seng Chua. 2023. Search-in-the-chain: Towards the accurate, credible and traceable content generation for complex knowledge-intensive tasks. *arXiv preprint arXiv:2304.14732*.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023a. 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*. ACM.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023b. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, page 188–202, New York, NY, USA. Association for Computing Machinery.

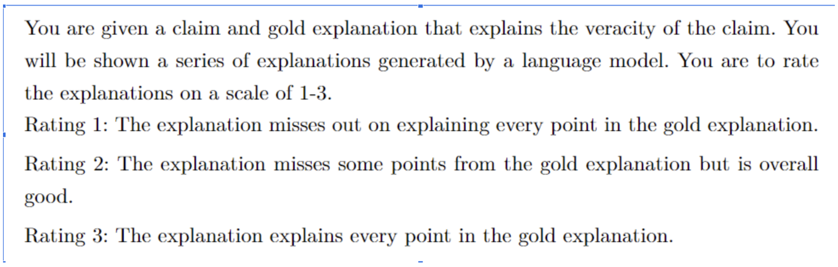
Fengzhu Zeng and Wei Gao. 2024. Justilm: Few-shot justification generation for explainable fact-checking of real-world claims. *arXiv preprint arXiv:2401.08026*.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *arXiv preprint arXiv:2310.00305*.

A Appendix

A.1 Instructions to Annotators

The instructions to annotators for the evaluation of the Explanation Generation Task is provided in Figure 5.



You are given a claim and gold explanation that explains the veracity of the claim. You will be shown a series of explanations generated by a language model. You are to rate the explanations on a scale of 1-3.

Rating 1: The explanation misses out on explaining every point in the gold explanation.

Rating 2: The explanation misses some points from the gold explanation but is overall good.

Rating 3: The explanation explains every point in the gold explanation.

Figure 5: Annotation Instructions

A.2 Explanation Generation by Veracity Label

In addition to the overall ratings for the Human Annotation for Explanation Generation, we also provide the ratings for specific classes. Figure 6 shows the human annotation ratings for the explanations of supported claims. Figure 7 shows the human annotation ratings for the explanations of refuted claims.

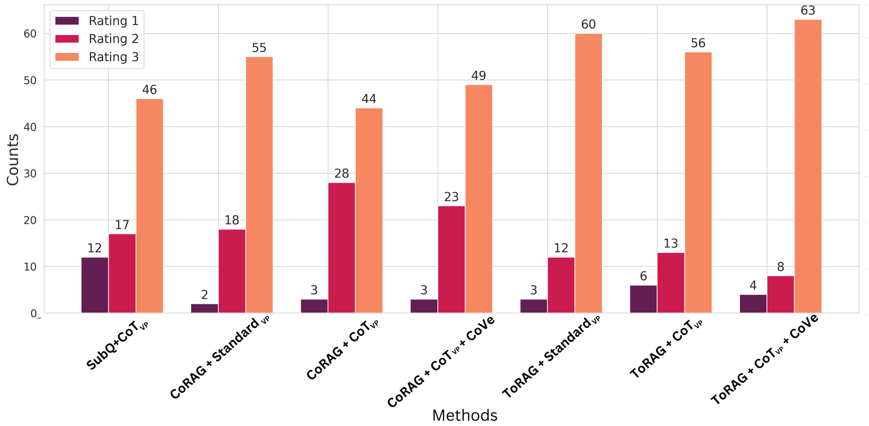


Figure 6: Annotator ratings for explanations of supported claims

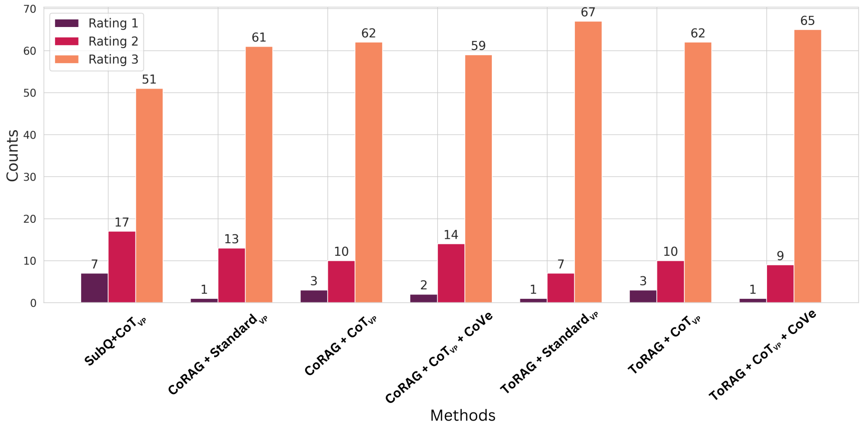
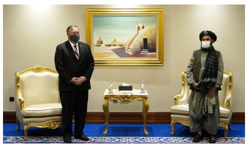
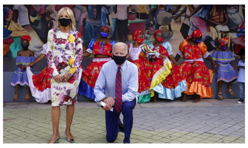


Figure 7: Annotator ratings for explanations of refuted claims

A.3 Discussing Multimodal RAG

We utilize reverse image search to extract captions of matching images from the web. We showcase the Image QA pairs for the examples in Table 2. The first example regarding Mike Pompeo showcases how GPT-4V is unable to identify the Afghan dignitary and the image context is unable to provide a name that could help fact-check the claim. However, using the image captions retrieved from the internet and prompting the evidence retrieval along with the image caption, GPT-4V is able to identify the Afghan dignitary as Mullah Abdul Ghani Baradar. The fact-check then continues to verify if Mullah Abdul Ghani Baradar was indeed ever the Afghan President or not. Similarly, in the third example with Joe Biden kneeling, the image captions extracted by reverse image search are able to add the additional information that Joe Biden was kneeling down to pose with dancers in Haiti. This information is crucial for the particular fact-check since it contextualizes the reason why Joe Biden was kneeling as well as detailing the event where the described act occurred.

Table 2: Example table with claims, images, and QA.

Claim	**Image**	**Generated Image Context**	**Image QA**
The man next to Mike Pompeo in a November 2020 photo is the guy the Trump administration helped get out of jail in 2018 and who is now the 'president' of Afghanistan.		The image shows Mike Pompeo standing next to a man wearing traditional Afghan attire and a face mask. The setting appears to be a formal meeting room with a painting on the wall, two gold-colored chairs, and a table with water bottles. Both individuals are standing apart, likely adhering to social distancing guidelines prevalent during the COVID-19 pandemic. The background and attire suggest the meeting is official in nature and likely related to diplomatic or governmental affairs. Claim stated in 2021.	Q. Who is the man standing next to Mike Pompeo in the November 2020 photo? A. Mullah Abdul Ghani Baradar
Says Joe Biden took a knee during the national anthem.		The image shows Joe Biden taking a knee, but it does not provide evidence that this act occurred during the national anthem. Without additional context, the claim cannot be confirmed solely based on this image. The time frame of the claim is 'Claim stated in 2020'.	Is there a specific date and location associated with the image of Joe Biden taking a knee? A. FILE - In this Monday, Oct. 5, 2020 file photo, Democratic presidential candidate former Vice President Joe Biden and his wife Jill Biden pose for a photo with dancers as they visit Little Haiti Cultural Complex in Miami.
The Trump administration worked to free 5,000 Taliban prisoners.		The image shows individuals, presumed to be Taliban prisoners, inside a bus with a guard standing nearby, which potentially correlates to the release of Taliban prisoners. The context suggests this may represent a prisoner release process.	Q. Were the individuals shown in the provided image actually Taliban prisoners being released as part of the agreement? A.'Taliban prisoners are released from Pul-e-Charkhi jail in Kabul, Afghanistan, Thursday, Aug. 13, 2020
These were not chemical irritants' used to clear a crowd. Pepper balls are 'not a chemical irritant.		The image shows law enforcement in protective gear amidst a haze that is consistent with the use of some form of crowd control substance, such as a chemical irritant. Visible smoke and the dispersing crowd strongly suggest the use of a substance to clear the area, counter to the claim that no chemical irritants were used. The presence of pepper balls would depend on identifying specific items or equipment in the scene that are known to dispense pepper balls.	Can we identify the specific equipment or methods used by law enforcement in the image to determine if pepper balls or another substance was deployed? A. ... The caption snippets suggest that teargas was used to clear Lafayette Park for a photo opportunity...

A.4 Example Questions

In Table 3 and Table 4 we show the questions asked by each approach.

Table 3: Example of Question-Answers asked by SubQ+CoT_{VP}, CoRAG and ToRAG

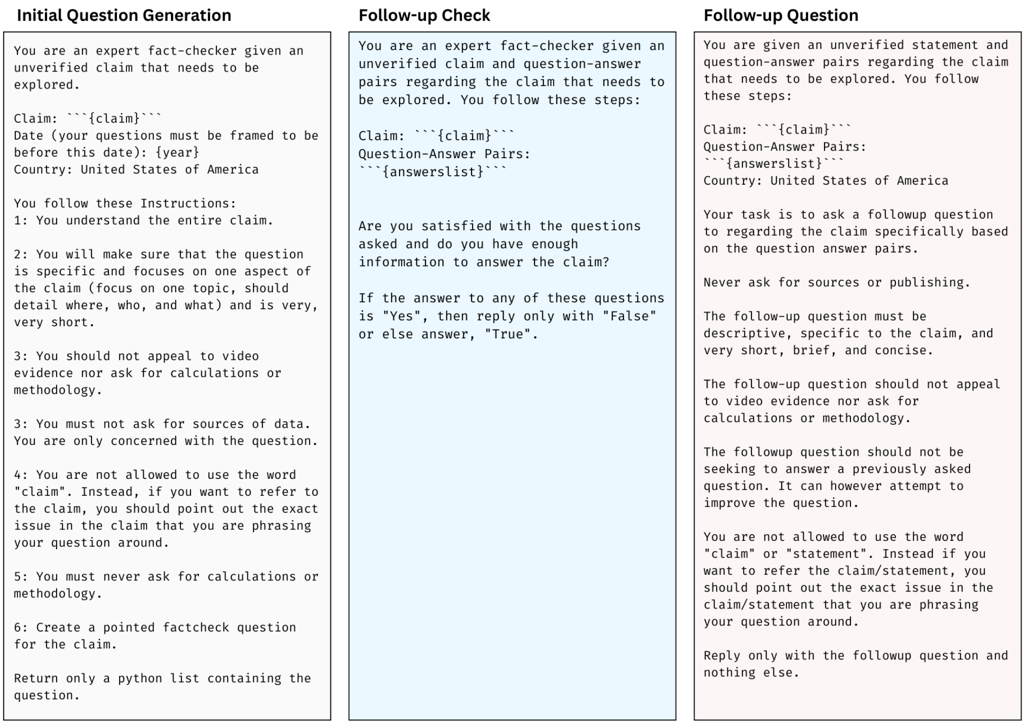
Claim: PPP on average provided a grant of around \$11,000 per employee	
SubQ+ CoT_{VP}	**Q. What was the total amount of funds disbursed through the Paycheck Protection Program (PPP) in 2020?** A.More than \$511 billion **Q. What criteria were used to determine the amount of grant each business received per employee under the PPP?"** A. ... determined based on the business's average monthly payroll costs... approximately 2.5 times the applicant's average monthly payroll... only up to \$100,000 of an employee's annual compensation was eligible for calculation. **Q. Were there any variations in the grant amount per employee based on the industry, location, or size of the business?** A. ... mentions different aspects of grants, such as administration costs, the impact of public money on grant receivers, and the employee retention credit, which suggests that grant amounts... **Q. How does the PPP define an "employee" for the purposes of calculating grants?** A. ...means individuals employed on a full-time, part-time, or other basis.
CoRAG	**Q. What was the average grant amount per employee provided by the Paycheck Protection Program in 2020?** A. ...for self-employed individuals and employees, the PPP loans could cover up to \$100,000 of annualized pay per employee for eight weeks, with a maximum of \$15,385 per individual." **Q. How does the maximum allowable amount of \$15,385 per individual under PPP loans relate to the average grant of around \$11,000 per employee?** A. The maximum allowable amount of \$15,385 per individual under PPP loans is derived from the limit set for self-employed individuals and employees, which is based on up to \$100,000 of annualized pay for eight weeks...On the other hand, the average grant of around \$11,000 per employee, as observed by Square Capital, reflects the actual average loan amount disbursed to small businesses through the PPP.
ToRAG	**Q. What was the typical grant amount per employee provided by the PPP as of mid-2020?** A. ... the average size of a PPP loan approved during the first wave was \$206,000, with 74% of the loans approved being for amounts of \$150,000 or less." **Q. What is the average monthly payroll cost per employee, as used in the PPP loan calculations, and how does this figure relate to the reported average grant of around \$11,000 per employee?** A. ...For a 24-week period, the maximum amount per individual is \$46,154, or for an eight-week period, it is \$15,385. **Q. Can you provide the total number of employees covered by PPP loans up to mid-2020 and the total amount of PPP funds disbursed by that time, to calculate the actual average grant per employee?** A. ... average grant per employee can be calculated as \$511 billion divided by 51 million employees, which equals approximately \$10,019.61 per employee.

Table 4: Example of Question-Answers asked by CoRAG and ToRAG

Claim: These 11 mass shootings involved an AR-15-style weapon.	
CoRAG	**Q.Were AR-15-style weapons used in 11 mass shootings in the United States before 2021?** A. . . . confirms that AR-15-style rifles were used in the 2012 Aurora, Colorado... **Q. Can you list the specific mass shootings, aside from the 2012 Aurora, Colorado incident, where AR-15-style rifles were reportedly used before 2021?** A. ... search results provided do not list specific mass shootings... stating that AR-15 rifles were used in 26 percent of the last 80 mass shootings...before 2021. **Q. Given the information that AR-15-style rifles were used in 26 percent of the last 80 mass shootings before 2021, identify other incidents where these rifles were reportedly used.** A. ... occurred when a gunman, armed with an AR-15-style rifle and body armor, killed nine people and wounded 27 others. **Q. Can you identify the specific characteristics or features that define an AR-15-style rifle as used in these incidents?** A. 1. Chambering . . . 2. Controls 3. Fast and easy reloading. . .

Claim: (Continued)	
ToRAG	**Q. What are the specific dates and locations of the 11 mass shootings mentioned as involving an AR-15-style weapon?** A. 1. Boulder, Colorado: March 22, 2021... 2. San Jose, California: May 26, 2021..." **Q. Were there any official reports or statements from law enforcement agencies confirming the use of AR-15-style weapons in the Boulder, Colorado, and San Jose, California shootings?** A. Yes, there were official reports confirming the use of an AR-15-style weapon in the Boulder, Colorado shooting. . . **Q. Can you provide details on the legal acquisition and ownership status of AR-15-style weapons by the shooters in the remaining nine mass shootings mentioned?** A. 1. Sutherland Springs church shooting: ... goods retailer violated the law ... 2. Boulder supermarket shooting: ... legally purchased the AR-15-style rifle ... 4. Pittsburgh synagogue shooting: ... like the AR-15 rifle used in the attack. 5. Las Vegas shooting: ... claim for selling AR-15s... 6. Orlando nightclub shooting: ... 7. San Bernardino shooting: ... 8. Sandy Hook Elementary School shooting: ..."

A.5 General Prompts in the RAGAR Approaches



Initial Question Generation

````
You are an expert fact-checker given an
unverified claim that needs to be
explored.

Claim: ```{claim}```
Date (your questions must be framed to be
before this date): {year}
Country: United States of America

You follow these Instructions:
1: You understand the entire claim.

2: You will make sure that the question
is specific and focuses on one aspect of
the claim (focus on one topic, should
detail where, who, and what) and is very,
very short.

3: You should not appeal to video
evidence nor ask for calculations or
methodology.

3: You must not ask for sources of data.
You are only concerned with the question.

4: You are not allowed to use the word
"claim". Instead, if you want to refer to
the claim, you should point out the exact
issue in the claim that you are phrasing
your question around.

5: You must never ask for calculations or
methodology.

6: Create a pointed factcheck question
for the claim.

Return only a python list containing the
question.
````

Follow-up Check

~~~
You are an expert fact-checker given an
unverified claim and question-answer
pairs regarding the claim that needs to
be explored. You follow these steps:

Claim: ```{claim}```
Question-Answer Pairs:
```{answerslist}```


Are you satisfied with the questions
asked and do you have enough
information to answer the claim?

If the answer to any of these questions
is "Yes", then reply only with "False"
or else answer, "True".
~~~

Follow-up Question

~~~
You are given an unverified statement and
question-answer pairs regarding the claim
that needs to be explored. You follow
these steps:

Claim: ```{claim}```
Question-Answer Pairs:
```{answerslist}```
Country: United States of America

Your task is to ask a followup question
to regarding the claim specifically based
on the question answer pairs.

Never ask for sources or publishing.

The follow-up question must be
descriptive, specific to the claim, and
very short, brief, and concise.

The follow-up question should not appeal
to video evidence nor ask for
calculations or methodology.

The followup question should not be
seeking to answer a previously asked
question. It can however attempt to
improve the question.

You are not allowed to use the word
"claim" or "statement". Instead if you
want to refer the claim/statement, you
should point out the exact issue in the
claim/statement that you are phrasing
your question around.

Reply only with the followup question and
nothing else.
~~~

Figure 8: Prompt for initial question-generation, Follow-up Check and Follow-up Question common to all RAGAR approaches

A.6 Prompts Specific to Tree of RAG

QA Elimination

```
You are an expert fact-checker. You are given a claim and a question-answer pair containing
3 questions alongwith their answers.
Claim: ```{claim}```
Question-Answer Pair: ```{qapairs}```

These questions and answers are seeking to help fact-check the claim. You as a fact-checker
have to pick only one of these question answer pairs. Here are your guidelines to pick:
- Pick the question-answer pair that is the most relevant towards answering the claim.
- Pick the question-answer pair that divulges the most information.
- Pick the question-answer pair that reveals new additional information.
- Do not pick the question-answer pair that is unsure with its answer and does not have the
answer.
- Pick the question answer pair that can answer the question precisely and is the most
detailed.

You have to pick the one that most matches these criteria.
Reply only with the question answer pair as a dictionary with question as key and answer as
value.
```

Figure 9: Prompt for QA Elimination

A.7 Prompts for Veracity Prediction

A.7.1 Standard Veracity Prediction Prompt

Veracity Prediction and Explanation

```
You are a well-informed and expert fact-checker.
You are provided with question-answer pairs regarding the following claim: {claim}

These are the provided questions and relevant answers to the question to verify the claim:
<{evidence_pairs}>

Based strictly on the main claim and the question-answers provided (ignoring questions regarding image if they
dont have an answer), You have to provide:

- claim: the original claim,

- rating: the rating for claim should be "supported" if and only if the Question Answer Pairs specifically
support the claim, "refuted" if and only if the Question Answer Pairs specifically refute the claim or "failed":
if there is not enough information to answer the claim appropriately.

- factcheck: and the detailed and elaborate fact-check paragraph.

please output your response in the demanded json format
```

Figure 10: Prompt for Standard Veracity prediction

A.7.2 Zero Shot Chain of Thought Veracity Prediction

Zero Shot Chain of Thought Veracity Prediction

```
You are a well-informed and expert fact-checker.

You are provided with question-answer pairs regarding the following claim: {claim}

Question-Answer Pairs:
<{evidence_pairs}>

Based strictly on the main claim, and the question-answers provided (ignoring questions regarding image if they
dont have an answer), you will provide:

rating: The rating for claim should be one of "supported" if and only if the Question Answer Pairs specifically
support the claim, "refuted" if and only if the Question Answer Pairs specifically refutes the claim or
"failed": if there is not enough information to answer the claim appropriately.

Is the claim: {claim} "supported", "refuted" or "failed" according to the available questions and answers?
Lets think step by step.
```

Figure 11: Prompt to get the CoT Veracity Prediction from the question-answer pairs and the claim

A.7.3 Chain of Verification Veracity Prediction

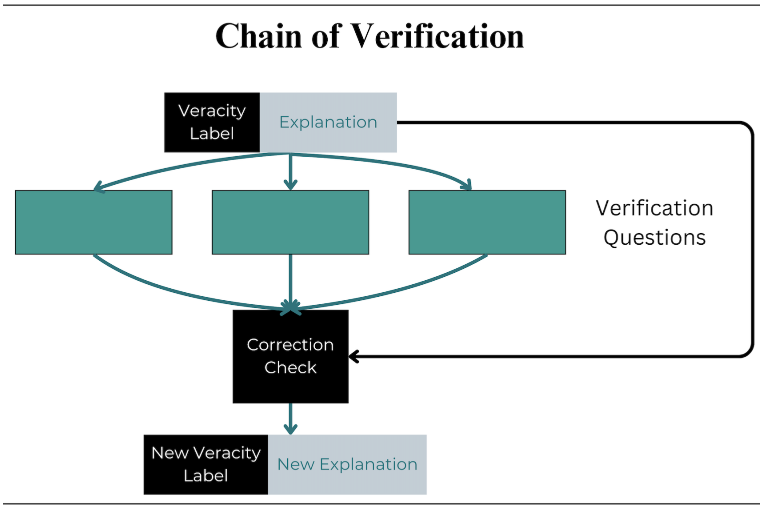


Figure 12: Pipeline of the CoVe Veracity Prediction

Verification Questions

```
You are a well-informed and expert fact-checker.
You are provided with the fact-check regarding the given claim and also the year the claim was made in.

Claim: {claim}
Fact-Checked Response: {factcheck}
Year the claim was made (specify it in your question): {year}

You are to generate verification questions.
A verification question is defined as a question that seeks to directly confirm whether a point made in the fact-checked response is true or false.

Your task is the following:
1. Read the entire fact-check.

2. Identify overall points mentioned in the factcheck.

3. Create pointed verification questions by rephrasing the point verbatim as a Yes/No question for the overall points mentioned in the fact-check.

4. The question must seek to gain answers in case of missing information suggested in the fact-check.

5. You must stick only to the overall points mentioned in the fact-check, do not create questions for unnecessary extra information.

Instruction: You are not allowed to use the word "claim" or "statement". Instead if you want to refer the claim/statement, you should point out the exact issue in the claim/statement that you are phrasing your question around.


Return only the pointed verification questions each seperated with a "~~~" symbol.
```

Figure 13: CoVe Verification Questions prompt

Correction Check

```
You are a well-informed and expert fact-checker.
You are provided with a factcheck and its correction qa pairs regarding the following claim: {claim}

Original FactCheck:
<{factcheck}>

Correction QA: {corrections}

Based strictly on the main claim, the original factcheck and the question-answers provided (ignoring questions regarding image if they dont have an answer), you will:

- If the corrections contain information that differs from the original factcheck, then create a new factcheck based on the corrected information and explain whether this changes the veracity of the original claim.

- If the corrections do not contain any new factcheck information, then simply return the original factcheck back.
```

Figure 14: CoVe Corrections Prompt