

ClaimCheck: Real-Time Automatic Fact-Checking with Small Language Models

Akshith Reddy Putta*, Jacob Devasier*, Chengkai Li

University of Texas at Arlington
{akshith.putta, jacob.devasier, cli}@uta.edu

Abstract

We introduce ClaimCheck, an LLM-guided automatic fact-checking system designed to verify real-world claims using live Web evidence and small language models. Unlike prior systems that rely on large, closed-source models and static knowledge stores, ClaimCheck employs a transparent, step-wise verification pipeline that mirrors human fact-checking workflows consisting of Web search query planning, Web-based evidence retrieval and summarization, evidence synthesis and re-retrieval, and claim verdict evaluation. Each module is optimized for small LLMs, allowing the system to deliver accurate and interpretable fact-checking with significantly lower computational requirements. Despite using a much smaller Qwen3-4B model, ClaimCheck achieves state-of-the-art accuracy of 76.4% on the AVeriTeC dataset, outperforming previous approaches using much larger models like GPT-4o and Llama 3.

Introduction

The proliferation of misinformation across social media has created an urgent need for accessible, reliable fact-checking tools. While manual fact-checking by professional organizations remains the gold standard, it is time-intensive and cannot scale to meet the volume of claims requiring verification. Recent automated fact-checking systems have achieved strong performance by combining retrieval-augmented generation with large language models (LLMs) (Rothermel, M. et al. 2024; Yoon, Y. et al. 2024; Malon 2024), but these approaches face significant barriers to widespread adoption: they typically depend on large, closed-source models that are computationally or monetarily prohibitive, or difficult to deploy (Schlichtkrull, M. et al. 2024; Braun et al. 2024).

In this work, we present ClaimCheck, an LLM-guided automatic fact-checking system that enables both experts and non-experts to verify real-world claims using real-time Web evidence and transparent, modular reasoning. Given a claim, the system displays the fact-checking process in real-time on a user-friendly interface and produces a step-by-step interpretable report that details how it planned searches, gathered and summarized evidence, synthesized information, and arrived at a final verdict.

*These authors contributed equally.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

A central objective of ClaimCheck is to democratize access to trustworthy fact-checking tools by demonstrating that effective automated fact-checking can be achieved with substantially smaller, more accessible models than previous systems. Smaller LLMs offer compelling advantages including reduced computational requirements, lower operational costs, and suitability for deployment on local or edge devices. However, such models commonly exhibit lower task-specific performance and limited reasoning capabilities compared to their larger counterparts. Previous efforts using smaller LLMs for fact-checking have encountered notable challenges. For instance, Putta, A. et al. (2025) using the Qwen2.5 7B model struggled with effectively synthesizing evidence and predicting the verdict for fact verification. Building on the modular approach established by DEFAME (Braun et al. 2024), this work demonstrates that these limitations can be systematically addressed through specialized modular design optimized for smaller LLMs that decomposes fact-checking into manageable components.

ClaimCheck’s workflow consists of five core stages: (1) query planning, (2) evidence retrieval, (3) evidence summarization/filtering, (4) evidence synthesis and re-retrieval, and (5) claim evaluation. These modules communicate via a centralized fact-checking report that serves as both an internal coordination mechanism and a user-facing explanation of the reasoning process. This design enables users to inspect the system’s decision-making at each step, addressing transparency requirements for fact-checking.

We evaluate ClaimCheck on the AVeriTeC dataset (Schlichtkrull, M. et al. 2024), a popular real-world factual claims dataset, applying strict temporal cutoffs to prevent data leakage and reflect realistic verification conditions. Despite using a much smaller model (Qwen3-4B), ClaimCheck achieves state-of-the-art accuracy (76.4%), outperforming prior approaches—PASS-FC (72.0%) (Zhuang 2025) and HerO (75.2%) (Yoon, Y. et al. 2024)—which rely on substantially larger models (e.g., LLaMA3.1 70B and GPT-4o) and use pre-fetched knowledge stores.

ClaimCheck

ClaimCheck is an LLM-guided automatic fact-checking system. It employs a modular architecture where each component performs a distinct function in the verification work-

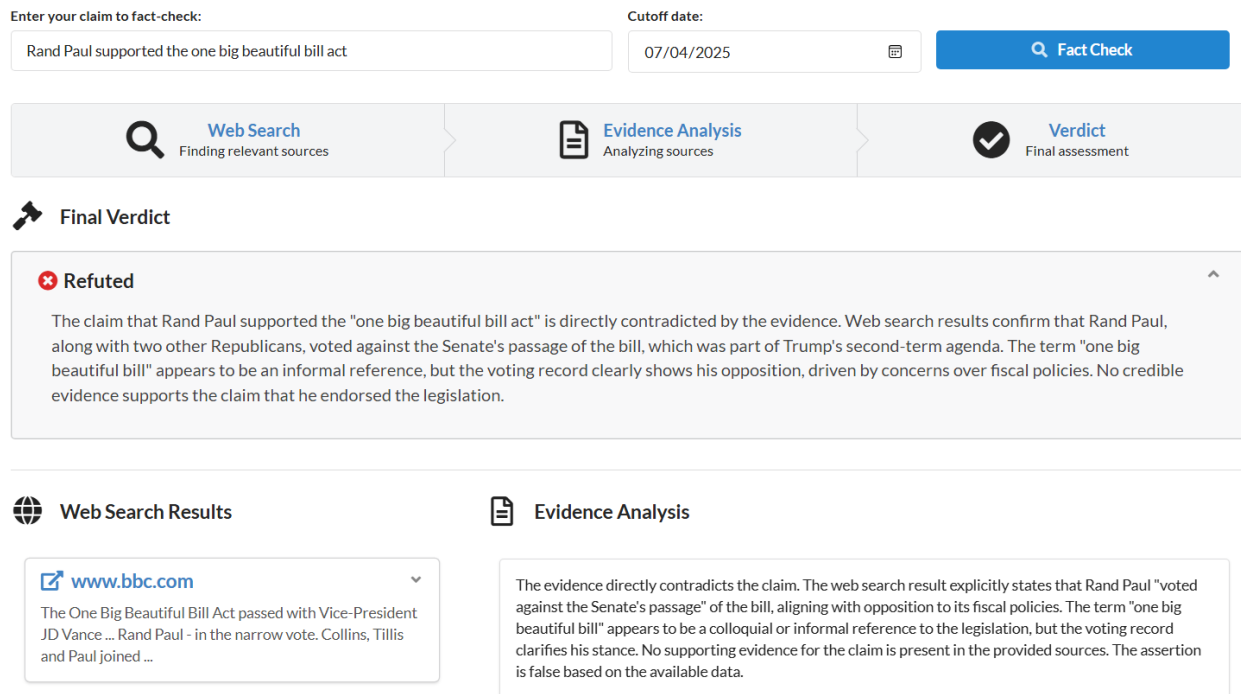


Figure 1: An example of our demo on the claim “Rand Paul supported the one big beautiful bill act.” A list of evidence articles is shown in the bottom left, a synthesized analysis of the evidence is shown in the bottom right, and the final verdict produced by our system is shown in the middle. Each module is populated as they are completed, with a progress bar shown at the top.

flow. We utilize Qwen3-4B (Qwen Team 2025) with thinking enabled across all modules as our LLM to ensure comprehensive reasoning at each step. The LLM prompts used for each module are presented in the appendix and are largely based on DEFAME’s (Braun et al. 2024) prompts.

System Architecture

The ClaimCheck system operates through a structured pipeline mirroring human fact-checking practices while leveraging the scalability of automated systems. The approach begins by analyzing the claim and constructing Web search queries to gather evidence (*planning*). Next, the system uses the queries to find and collect relevant articles (*execution*) and summarizes key points about the claim from the articles (*evidence summarization*). Then, all of the evidence is synthesized to a single cohesive analysis to determine if additional information is needed (*evidence synthesis*) and collect it if so (*execution*). If no further evidence is needed, the system proceeds to fact verification. Finally, the system uses the synthesized evidence to determine the veracity of the claim and provide a justification (*evaluation*).

Fact-Checking Report. ClaimCheck generates a comprehensive fact-checking report for each input claim. This report serves as a central artifact that facilitates communication between modules by recording their intermediate outputs in a structured format. This enables later modules to reference earlier outputs and ensures consistent, interpretable outputs for users. Once the fact-checking process is com-

plete, users can download the report to review the system’s internal reasoning in each module.

Modules

Planning. The *planning* module utilizes the LLM to understand the input claim and construct a set of Web search queries to collect evidence to fact-check the claim. The LLM is instructed (Listing 1 in Appendix) to consider different aspects of the claim in case a claim may have multiple sub-claims. While we do not explicitly instruct the LLM to perform claim decomposition, as is common in many previous works (Braun et al. 2024; Iqbal, H. et al. 2024), Qwen3 often breaks down the claim in its thinking.

Execution. The *execution* module carries out the Web search using the queries created in the *planning* module. We utilize Serper (<https://serper.dev/>) for executing our queries using the Google search engine, returning URLs and snippets for each article. In our demo, we limit this to only the top-3 results to significantly increase the efficiency of the system. We present this list of URLs and snippets in the Web search results.

Evidence Summarization. The *evidence summarization* module uses the LLM (Listing 2) to individually extract and summarize the relevant information from each collected article and discard articles which are not helpful for fact-checking the claim. Figure 2 shows how the evidence summary for each collected article is displayed to the user.



Figure 2: An example of a summarized article collected using ClaimCheck’s Web evidence search for the claim “Rand Paul supported the one big beautiful bill act.”

Evidence Synthesis. The *evidence synthesis* module uses the LLM (Listing 3) to synthesize all of the summarized evidence into a single coherent analysis that consolidates supporting and contradicting information, identifies patterns across sources, highlights key factual information with their supporting evidence, and assesses the overall reliability and consistency of the collected information. After the analysis is produced, the LLM will determine whether additional Web search is needed, taking into account any gaps in evidence and previous queries, and produce a Web search query if so. This will trigger the system to rerun the *execution* and *evidence summarization* modules.

Evaluation. The *evaluation* module uses the LLM (Listing 4) to evaluate the output of the synthesized evidence and assigns a final verdict to the claim along with an explanation of the predicted veracity.

Experiments

Dataset and Evaluation AVeriTeC (Schlichtkrull, Guo, and Vlachos 2023) is a benchmark dataset for real-world automatic fact-checking. This dataset contains 4,568 authentic claims sourced from 50 different fact-checking organizations. Each claim is annotated with (1) one of four verdicts: Supported, Refuted, Conflicting Evidence/Cherry-picking, or Not Enough Evidence, (2) Question-answer pairs grounded in Web evidence which justify the verdict on the claim, (3) textual justifications which summarize how the collected evidence supports the verdict, and (4) metadata about the speaker, publication date, and location of the claim. These annotations are separated into Train (3,068 samples), Development (500 samples), and Test (1000 samples) parti-

Framework	Accuracy	LLM
⊗ GPT-4o mini	46.8%	GPT-4o mini
⊗ Qwen3 4B	52.0%	Qwen3-4B
⊗ GPT-4o	53.2%	GPT-4o
🗄 InFact	72.4%	GPT-4o
🗄 HerO	75.2%	Llama3.1 70B
🌐 Papelo	41.5%	GPT-4o mini
🌐 ClaimCheck*	67.0%	OpenAI o4-mini
🌐 GPT-4o mini	70.4%	GPT-4o mini
🌐 DEFAME	70.5%	GPT-4o
🌐 PASS-FC	72.0%	GPT-4o
🌐 GPT-4o	74.6%	GPT-4o
🌐 ClaimCheck*	76.0%	Qwen3-32B
🌐 ClaimCheck	76.4%	Qwen3-4B

* Results only evaluated on a subset of 100 claims.

Table 1: Verdict prediction accuracy of different frameworks on the AVeriTeC development dataset. 🌐 indicates using Web search for evidence, 🗄 indicates using the knowledge store, and ⊗ indicates using no evidence.

tions. The authors also released a knowledge store containing roughly 1,000 Web articles for each claim to support of-line retrieval systems.

We evaluate ClaimCheck on the development split of the AVeriTeC dataset, as the test set lacks gold verdict labels. The development set has a similar class distribution to other sets: 24.4% Supported, 61.0% Refuted, 7.6% Conflicting Evidence/Cherry-picking, and 7.0% Not Enough Evidence. To prevent temporal data leakage during Web retrieval, we restrict our search to documents published before the claim’s annotated publication date.

We report veracity prediction accuracy, defined as the proportion of claims for which the system correctly predicts the verdict. Our system is compared against the top-performing submissions from the 2024 AVeriTeC shared task (Schlichtkrull, M. et al. 2024): InFact (Rothermel, M. et al. 2024), Papelo (Malon 2024), and HerO (Yoon, Y. et al. 2024), as well as two recent fact-checking systems: DEFAME (Braun et al. 2024) and PASS-FC (Zhuang 2025). We also include additional baselines using GPT-4o and GPT-4o mini with OpenAI’s Web search tool (see appendix).

Table 1 presents the verdict prediction accuracy of ClaimCheck and other systems, grouped by evidence source. ClaimCheck achieves state-of-the-art performance with 76.4% accuracy using Qwen3-4B, outperforming all prior systems including those using significantly larger models.

Hybrid Thinking Ablation Study A key innovation in ClaimCheck is the use of Qwen3’s hybrid thinking capability. We conduct an ablation study on the previous 100 randomly sampled claims from the AVeriTeC development set to understand the contribution of thinking in each module. Table 2 summarizes the results of the experiment.

Configuration	Accuracy
All modules Think (baseline)	75.0%
⊘ Planning	71.0%
⊘ Evidence Summarization	66.0%
⊘ Evidence Synthesis	72.0%
⊘ Evaluation	65.0%
All modules No-Think (baseline)	54.0%
⊘ Planning	59.0%
⊘ Evidence Summarization	60.0%
⊘ Evidence Synthesis	60.0%
⊘ Evaluation	61.0%

Table 2: Ablation study of reasoning module configurations on the random subset of 100 claims from AVeriTeC’s dev set. Each row modifies the *think* setting of a single module while others remain fixed. ⊘ indicates that thinking is enabled only for a particular module, and vice versa with ⊘.

The fully thinking configuration achieves 75.0% accuracy, outperforming the all no-think baseline (54% accuracy). This 21% improvement demonstrates the significant value of reasoning capabilities in fact-checking tasks. When disabling thinking in individual modules from the all-think baseline, we observe varying performance impacts. The *evidence summarization* and *evaluation* modules show the largest degradation (9–10 percentage points), indicating these components are most critical for effective reasoning. The *planning* and *evidence synthesis* modules show smaller reductions (3–4 percentage points), suggesting these tasks are less reasoning-intensive.

Starting from the all no-think baseline, enabling thinking in any single module provides consistent but modest improvements (5–7 percentage points), with no single module dominating. This suggests that while each module benefits from thinking, the full pipeline requires comprehensive reasoning for optimal performance. Based on these findings, we enable thinking across all modules in the final system.

Model Scale Analysis We investigate whether larger models improve performance by comparing the Qwen3-4B and Qwen3-32B variants of ClaimCheck. Surprisingly, we observe no significant performance difference between the two model sizes (76.4% vs. 76.0% accuracy, as shown in Table 1), suggesting that model scale alone is not the primary driver of performance in our pipeline. This finding supports our hypothesis that effective fact-checking systems benefit more from high-quality evidence retrieval and careful architectural design than from raw model capacity. However, when using OpenAI’s o4-mini as the base LLM, accuracy drops significantly to 67.0%. This degradation likely stems from our prompting strategy being optimized for Qwen3 models, highlighting the importance of model-specific optimization in modular architectures.

Conclusion

This work presents ClaimCheck, an LLM-guided fact-checking system that verifies real-world claims using live

Web evidence and small, accessible models. By decomposing fact-checking into discrete, interpretable stages, ClaimCheck enables effective reasoning with the Qwen3-4B model, outperforming systems powered by much larger LLMs on the AVeriTeC dataset. Our results show that thoughtful architectural design and specialized prompting can overcome limitations typically associated with smaller LLMs. ClaimCheck offers a promising direction for democratizing access to trustworthy, transparent fact verification.

References

- Braun, T.; Rothermel, M.; Rohrbach, M.; and Rohrbach, A. 2024. DEFAME: Dynamic Evidence-based FAct-checking with Multimodal Experts. *arXiv preprint arXiv:2412.10510*.
- Iqbal, H. et al. 2024. OpenFactCheck: A Unified Framework for Factuality Evaluation of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 219–229. Miami, Florida, USA: Association for Computational Linguistics.
- Malon, C. 2024. Multi-hop Evidence Pursuit Meets the Web: Team Papelo at FEVER 2024. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, 27–36. Miami, Florida, USA: Association for Computational Linguistics.
- Putta, A. et al. 2025. ClaimCheck: Automatic Fact-Checking of Textual Claims using Web Evidence. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, 303–316. Albuquerque, New Mexico, USA: Association for Computational Linguistics. ISBN 979-8-89176-229-9.
- Qwen Team. 2025. Qwen3 Technical Report. arXiv:2505.09388.
- Rothermel, M. et al. 2024. InFact: A Strong Baseline for Automated Fact-Checking. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, 108–112. Miami, Florida, USA: Association for Computational Linguistics.
- Schlichtkrull, M. S.; Guo, Z.; and Vlachos, A. 2023. AVeriTeC: A Dataset for Real-world Claim Verification with Evidence from the Web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Schlichtkrull, M. et al. 2024. The Automated Verification of Textual Claims (AVeriTeC) Shared Task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, 1–26. Miami, Florida, USA: Association for Computational Linguistics.
- Yoon, Y. et al. 2024. HerO at AVeriTeC: The Herd of Open Large Language Models for Verifying Real-World Claims. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, 130–136. Miami, Florida, USA: Association for Computational Linguistics.
- Zhuang, Z. 2025. PASS-FC: Progressive and Adaptive Search Scheme for Fact Checking of Comprehensive Claims. arXiv:2504.09866.

Paper Checklist to be included in your paper

1. For most authors...

- (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? *NA*
- (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? *Yes, to the best of our knowledge.*
- (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? *Yes, including prompts used.*
- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? *NA*
- (e) Did you describe the limitations of your work? *No, due to limited available space*
- (f) Did you discuss any potential negative societal impacts of your work? *Yes, in the ethics and risks section of the appendix.*
- (g) Did you discuss any potential misuse of your work? *Yes, in the ethics and risks section of the appendix.*
- (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? *Yes, in the appendix.*
- (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? *Yes.*

2. Additionally, if your study involves hypotheses testing...

- (a) Did you clearly state the assumptions underlying all theoretical results? *NA*
- (b) Have you provided justifications for all theoretical results? *NA*
- (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? *NA*
- (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? *NA*
- (e) Did you address potential biases or limitations in your theoretical framework? *NA*
- (f) Have you related your theoretical results to the existing literature in social science? *NA*
- (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? *NA*

3. Additionally, if you are including theoretical proofs...

- (a) Did you state the full set of assumptions of all theoretical results? *NA*
- (b) Did you include complete proofs of all theoretical results? *NA*

4. Additionally, if you ran machine learning experiments...

- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? *No, but the code for the demo is available at <https://anonymous.4open.science/r/claimcheck-8FDB/>.*
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? *NA*
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? *NA*
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? *Yes, in the appendix.*
 - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? *Yes, but it is limited due to space constraints.*
 - (f) Do you discuss what is “the cost” of misclassification and fault (in)tolerance? *Yes, in the ethics and risks section.*
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
- (a) If your work uses existing assets, did you cite the creators? *Yes.*
 - (b) Did you mention the license of the assets? *NA*
 - (c) Did you include any new assets in the supplemental material or as a URL? *Yes, the URL to the demo.*
 - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? *NA*
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? *NA*
 - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? *NA*
 - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? *NA*
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
- (a) Did you include the full text of instructions given to participants and screenshots? *NA*
 - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? *NA*
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? *NA*
 - (d) Did you discuss how data is stored, shared, and de-identified? *NA*

Reproducibility

LLM Hyperparameters. ClaimCheck uses LLMs without finetuning for any tasks, with temperature

set to 0.6 and top- p to 0.95 for all Qwen3 models, and temperature set to 0.3 and top- p to 1.0 for o4-mini. We used Ollama (<https://ollama.com/>) to run Qwen3 models and the OpenAI API to run o4-mini. We evaluated our system using 1x Nvidia H100. For the Web search-enabled GPT-4o/GPT-4o mini, we used `gpt-4o-search-preview-2025-03-11` and `gpt-4o-mini-search-preview-2025-03-11`.

GitHub Copilot was used in the development of some of the code in this system. Generative AI tools were also used in early versions of the manuscript to improve writing and flow.

Ethics and Risks

Misinformation Amplification. While ClaimCheck is designed to detect and mitigate misinformation, its reliance on live Web search introduces potential risks. Specifically, the system may inadvertently retrieve and summarize low-quality or misleading sources, especially if those are ranked highly by search engines. Although our modular pipeline includes filtering and synthesis stages to reduce this risk, the system cannot guarantee that all retrieved content is accurate or representative. To mitigate this, we provide transparent reporting so users can trace each verdict back to its evidence.

Over-reliance and Misuse. Users may overestimate the capabilities or correctness of the system, especially given its step-by-step justifications and high reported accuracy. However, the system is not infallible and should not be used as a substitute for human judgment in high-stakes domains such as legal decisions, medical claims, or policy debates. To discourage misuse, we explicitly label the system as experimental and provide downloadable reports to encourage further human review.

Bias and Fairness. The LLMs powering ClaimCheck are pretrained on Web-scale data and may reflect underlying social, political, or cultural biases. These biases can surface in query planning, evidence summarization, and especially in the final evaluation stage. Although we use a relatively small and accessible model (Qwen3-4B), its reasoning remains influenced by the distributions in its training data. We partially address this through modular decomposition and prompt engineering, but biases in evidence selection or synthesis remain a concern—particularly for politically sensitive or underrepresented topics.

Prompts

We adopt the prompting structure from DEFAME (Braun et al. 2024) and adapt it to better suit smaller models such as Qwen3-4B. Specifically, we simplify instructions, reduce verbosity, and structure prompts to minimize context window overflow, which smaller LLMs are more susceptible to.

The prompts are depicted in Listings 1– 4. Particularly, the following verdict descriptions are embedded in the “Rules” section of the evaluation prompt (Listing 4).

Supported - The claim is directly and clearly backed by strong, credible evidence. Minor uncertainty or lack of detail does not disqualify a claim from being Supported if the

main point is well-evidenced. - Use Supported if the overall weight of evidence points to the claim being true, even if there are minor caveats or not every detail is confirmed.

Refuted - The claim is contradicted by strong, credible evidence, or is shown to be fabricated, deceptive, or false in its main point. - Use Refuted if the central elements of the claim are disproven, even if some minor details are unclear. - Lack of any credible sources supporting the claim does not mean “Not Enough Evidence”—it means the claim is Refuted.

Conflicting Evidence/Cherry-picking - Only use this if there are reputable sources that directly and irreconcilably contradict each other about the main point of the claim, and no clear resolution is possible after careful analysis. - Do NOT use this for minor disagreements, incomplete evidence, or if most evidence points one way but a few sources disagree.

Not Enough Evidence - Only use this if there is genuinely no relevant evidence available after a thorough search, or if the claim is too vague or ambiguous to evaluate. - Do NOT use this if there is some evidence, even if it is weak, or if the claim is mostly clear but not every detail is confirmed. - This is a last-resort option only.

Listing 1: Prompt for Planning

The available knowledge is insufficient to assess the Claim. Therefore, propose a set of actions to retrieve new and helpful evidence. Adhere to the following rules:

- The actions available are listed under Valid Actions, including a short description for each action. No other actions are possible at this moment.
- For each action, use the formatting as specified in Valid Actions.
- Include all actions in a single Markdown code block at the end of your answer.
- Propose as few actions as possible but as much as needed. Do not propose similar or previously used actions.
- Consider Both Modalities Equally: Avoid focusing too much on one modality at the expense of the other, but always check whether the text claim is true or false.
- Compare Image and Caption: Verify the context of the image and caption.

Valid Actions:

web_search: Run an open web search for related webpages.

Examples:

```
web_search("New Zealand Food Bill 2020")
```

Record:

```
{record}
```

Claim: {claim}

Your Actions:

Listing 2: Prompt for Evidence Summarization

In order to find evidence that helps your fact-check, you just ran a web search, which yielded a Search Result. Your task right now is to summarize the Search Result concisely in at most 5 sentences, only including information that is relevant to the Claim you are checking.

What to include:

- Information that might be useful for fact-checking the claim (see Record).
- If available: the release date as well as the author or the publisher (e.g., the media company) of the search result.

Do NOT include:

- Advertisements.
- Any other information unrelated to the Record or the Claim.

Additional Rules:

- Do not add any additional information besides the information in the Search Result. Also, do not add any information that is not related to the claim, even if it is mentioned in the Search Result.
- If the Search Result doesn't contain any relevant information for the fact-checking work, print only one word in capital letters, do not include anything else: NONE.
- Keep your writing style consistent with the provided Examples.
- Try to filter out relevant information even if the Search Result is in a different language.

Claim: {claim}

Evidence:

```
{url}  
{search_result}
```

Record:

```
{record}
```

Your Summary:

Listing 3: Prompt for Evidence Synthesis

Instructions

You just retrieved new Evidence. Now, analyze the Claim's veracity using the evidence. Always adhere to the following rules:

- Focus on developing new insights. Do not repeat larger parts from the Record. Do not restate the Claim.
- Write down your thoughts step-by-step. Whenever necessary, you may elaborate in more detail.
- Depending on the topic's complexity, invest one to three paragraphs. The fewer, the better.
- If you find that there is insufficient information to verify the Claim, explicitly state what information is missing.
- If you cite web sources, always refer to them by including their URL as a Markdown hyperlink.
- Use information only from the recorded evidence:
- Avoid inserting information that is not implied by the evidence. You may use commonsense knowledge, though.

If it is extremely necessary to retrieve more evidence, you can propose actions to the user. If not necessary, do not add anything else other than the reasoning.

Adhere to the following rules:

- The actions available are listed under Valid Actions, including a short description for each action. No other actions are possible at this moment.
- For each action, use the formatting as specified in Valid Actions.
- Propose as few actions as possible but as much as needed. Do not propose similar or previously used actions.
- Include all actions in a single Markdown code block at the end of your answer.

Valid Actions:

web search: Run an open web search for related webpages.

Examples:

```
web_search("New Zealand Food Bill 2020")
```

Record:

```
{record}
```

Your Analysis:

Listing 4: Prompt for Evaluation

Instructions

Determine the Claim's veracity by following these steps:

1. Briefly summarize the key insights from the fact-check (see Record) in at most one paragraph.
2. Write one paragraph about which one of the Decision Options applies best. Include the most appropriate decision option at the end and enclose it in backticks like `this`.

Decision Options:

Supported|Refuted|Conflicting Evidence/Cherry picking|Not Enough Evidence

Rules:

```
{verdict_descriptions}
```

Record:

```
{record}
```

Your Judgement: