Fact-checking AI-generated news reports: Can LLMs catch their own lies?

Anonymous ACL submission

Abstract

In this paper, we evaluate the ability of Large Language Models (LLMs) to assess the veracity of claims in "news reports" generated by themselves or other LLMs. Our goal is to determine whether LLMs can effectively fact-check their own content, and identify limitations that future research needs to focus on to build systems to fact-check LLM-generated content. Our findings indicate that LLMs are more effective at assessing claims in national or international news stories than in local news stories, better at evaluating static information than dynamic information, and better at verifying true claims compared to false ones. We hypothesize that this disparity arises because the former types of claims are better represented in the training data. Additionally, we find that incorporating search engine results in a Retrieval-Augmented Generation (RAG) setup significantly reduces the number of unassessable claims. However, it also increases incorrect assessments, due to both irrelevant retrievals and LLM reasoning errors. This diagnostic evaluation highlights the need for future research on fact-checking machine-generated content to prioritize (i) improving the precision and relevance of retrieved information, (ii) improving the reasoning abilities of LLMs, and (ii) building human-in-the-loop systems when no supporting information can be found.

1 Introduction

Large Language Models (LLMs) have revolutionized the field of Natural Language Processing (NLP), effortlessly performing tasks that were traditionally considered highly challenging. Their performance is particularly impressive in generating natural language text. Models like GPT-4 can generate coherent, fluent summaries, accurately translate text between languages (especially those with a strong online presence and ample training data), and refine human writing to enhance fluency and appropriateness in tone and style for specific purposes. This technology has the potential to significantly increase productivity across many industries, offering endless applications. However, with this potential also come huge risks if they are not used properly. One of the main risks is that they can be easily used to generate convincing and yet factually incorrect text, either intentionally or unintentionally. For example, with a simple prompt like "Generate a news report about volcano eruption in Massachusetts, USA", GPT-4 would happily generate a news report starting with this paragraph:

> "Massachusetts, USA – May 7, 2025 – In an unprecedented and shocking event, a volcanic eruption has occurred in the state of Massachusetts, an area not typically associated with volcanic activity. The eruption took place early this morning in the central part of the state, near the town of Worcester, sending residents and scientists alike into a state of disbelief and concern."

Although there has never been a volcanic eruption in reality, the news report is coherent and fluent. Coupled with modern media platforms, such LLM-generated content can quickly spread and reach a large audience. An example is the emergence of AI "news" farms that produce news reports with LLMs to generate advertising revenue with little concern for their impact on society (Puccetti et al., 2024). The machine-generated reports can cause confusion and chaos and disrupt the proper functioning of the society.

In this study, we present experimental results to assess whether LLMs can identify the truthfulness of the news stories they generate and detect factual errors within them, aiming to uncover limitations that can guide future research on fact-checking LLM-generated content.

We generated 182 news stories with a simple prompt such as "Write a story about Kobe Bryant rejoining the Lakers" with two LLMs, GPT-4o (OpenAI et al., 2024) and GLM (Du et al., 2022), all stories with some incorrect claims. The false stories vary in degree of falsehood. Some describe impossible events, like Kobe Bryant rejoining the Lakers. Others depict highly improbable scenarios, such as a volcano erupting in Massachusetts. Still others distort real events by misreporting the time, location, or participants.

We conducted experiments in two settings. In the first, we input the full story into GPT-4o and GLM and asked whether it was truthful. In the second, we manually broke each story into checkable atomic claims—specific events, factual states (e.g., "Massachusetts borders New Hampshire"), or recurring events not tied to a specific time. We then decontextualized these claims (Choi et al., 2021) to enable verification outside the original story. In this setting, we also tested a Retrieval-Augmented Generation (RAG) approach by querying each claim via the Google Serper API[1] and providing the retrieved results to GPT-4 for fact-checking (Lewis et al., 2020).

Our results show that GPT-4o and GLM perform well at flagging clearly false stories involving well-known entities (e.g., Kobe Bryant rejoining the Lakers), but struggle with recent but unlikely events. At the atomic claim level, a substantial number of claims are misclassified—either factually correct claims are judged false, or false claims are judged true. An even larger portion of claims receive no definitive judgment, highlighting uncertainty in the models' assessments.

Using search results retrieved via the Google Serper API significantly reduces the number of unassessable claims, but also leads to more correct and incorrect assessments. Errors stem from irrelevant retrievals or reasoning flaws in the LLMs. Even with RAG, a notable portion of claims remain unassessable, particularly those involving novel events not covered by existing knowledge sources. While recent work has shown that external tools can improve LLM factuality (Gou et al., 2023), such methods are limited for fact-checking machine-generated news, which often includes unverifiable or emerging information. Addressing these challenges may require new human-in-the-loop approaches tailored to novel event detection and verification, in addition to improving search relevance and the reasoning ability of LLMs.

[1]https://serper.dev/

The rest of the paper is organized as follows. In Section 2, we discuss related work. In Section 3, we present our method for generating news stories, extract "atomic" claims, using LLMs to assess the veracity of these stories and claims, and manually verifying the assessments performed by LLMs themselves. We present experimental results in Section 4, and discuss these results in Section 5. We conclude in Section 6.

2 Related Work

Dataset statistics and comparison. Our data set is most similar to FactScore (Min et al., 2023) in that both consist of long-form texts generated by LLMs; however, two key differences separate our data sets. First, while FactScore focuses on biographies of Wikipedia entities, our dataset consists of LLM-generated news reports that include time-sensitive content, making them inherently harder to fact check due to the absence of a preexisting knowledge source. Second, while every short sentence in the biographies of FactScore is treated as an independent factual claim, our news reports often contain vague or subjective content, necessitating manual extraction of only those claims that are verifiably checkable. The following data sets are also broadly related to ours, but there are significant differences. Datasets like PROPANEWS (arXiv:2203.05386) are created by replacing sentences in real news articles with plausible but fake content to mimic factual claims made by humans. The FEVER data set consists of individual claims verified by Wikipedia. The EX-FEVER data set is also based on Wikipedia but requires multi-hop reasoning to fact-check to enhance explainability. The AVeriTeC data set contains real-world claims that can be checked against web sources. All these data sets are collections of individual claims created with the assumption that there is a knowledge source against which these claims can be verified.

Fact-checking human or machine-generated content. There is an active NLP research community focused on developing automatic methods to fact-check false claims, such as those made by politicians (Nakov et al., 2021; Deng et al., 2024; Yuan and Vlachos, 2024; Schlichtkrull et al., 2024). There is also more recent work on fact-checking machine-generated content (Min et al., 2023; Wang et al., 2024; Fadeeva et al., 2024). Previous work on fact-checking false claims made by either humans or machines typically assumes there is an information source, usually a published source on the Internet, against which the claims can be checked. However, events reported in machine-generated news stories that we are interested in, such as the volcano eruption example, are often assumed to be new occurrences that cannot be cross-verified against any existing public sources, although they may still contain claims about the real world that can be fact-checked. This poses novel challenges that are not present in people biographies used in previous studies (Min et al., 2023; Fadeeva et al., 2024).

3 Method

Our experiment on fact-checking LLM-generated news stories consists of four steps. First, we use two LLMs to generate a set of news stories with varying levels of factual inaccuracy. Next, from these stories, we manually extract verifiable atomic claims and decontextualize them, creating standalone claims that can be verified independently of the original story. In the third step, we prompt each LLM to evaluate the veracity of news stories generated by itself or the other LLM, as well as to assess the individual atomic claims. Finally, we conduct a human evaluation to determine the accuracy of the LLMs' veracity assessments.

3.1 News Report Generation with LLMs

To evaluate the claim verification capabilities of GPT and GLM, we first prompt both models to generate a set of 182 news articles, including 92 news artiles generated by GPT-4o and 90 articles generated by GLM. Each prompt is designed around scenario-based inputs that intentionally contain factual inconsistencies. The following is an example prompt that contains a time error, as the time of 2024 Australian Open women's final is January 27, not January 20:

> "Generate a news report about Aryna Sabalenka winning the 2024 Australian Open Women's final, held at Rod Laver Arena on January 20, as Aryna Sabalenka beat Zheng Qinwen (6-3, 6-2)."

All these inconsistencies are designed around four critical aspects of a scenario: the event itself, along with its time, location, and participants. To rigorously test the models' understanding of both nationally recognized and locally relevant information, we control the scope of the generated content by introducing both local and national news categories. The distinction between these categories serves as a critical factor in our evaluation, allowing us to evaluate how effectively each model handles claims involving specific local information versus those based on widely known national knowledge. This is motivated by prior research suggesting that LLMs may have greater exposure to widely discussed national or international events, given the nature of the large, diverse datasets they are trained on (Kandpal et al., 2023). When generating the news stories, we ensure that the same general template is used for all prompts, varying only the scenarios for each different story. By using consistent prompts, we ensure that differences in model performance can be attributed to the model's capabilities rather than variability in the inputs. This approach allows us to build a diverse and representative dataset that rigorously tests each LLM's ability to identify and evaluate issues across different aspects of the generated content.

3.2 Manual claim Extraction

After generating the news reports, we manually extracted all checkable claims from the GPT-generated content. Each claim is a clear, verifiable statement with specific details such as time, location, participants, or events. We adhered to criteria that required each checkable claim to contain precise, unambiguous information—such as exact dates, locations, or identifiable participants. Vague or generic statements, like "Sabalenka had a great match" were excluded, as they lack objective, verifiable details. This approach ensured that only claims containing concrete, factual information were selected for manual extraction. We manually decontextualize claims by resolving pronominal and other anaphoric expressions, and by supplementing events with time, location, and participant details when they are clear from the context, ensuring that each claim is independently verifiable.

The following are example claims illustrating various types of factual inaccuracies:

- **Time error:** "Aryna Sabalenka triumphed over Zheng Qinwen to win the 2024 Australian Open Women's final at Rod Laver Arena on January 20, 2024."
- **Location error:** "Aryna Sabalenka played against Zheng Qinwen in the 2024 Australian

Open Women's final at Margaret Court Arena on January 27."

- **Event error:** "In the third set of the 2024 Australian Open Women's Final at Rod Laver Arena on January 27, Zheng Qinwen broke Aryna Sabalenka's serve at 5-5 and won the set 7-5 to clinch the championship."
- **Participant and location error:** "Naomi Osaka and Iga Swiatek are battling for the prestigious Grand Slam title at the 2024 Australian Open Women's Final at the Margaret Court Arena on January 27, 2024."

Each article typically yields between 10-20 checkable decontextualized claims, depending on its length and complexity. This process ensures that the claims include all the necessary contextual information required for verification, maintaining the integrity and relevance of the claims within the broader context of the news reports. From the 182 articles we have extracted 2,323 total atomic claims, including 1186 claims from the 92 news reports generated by GPT-4o, and 1137 claims from the 90 reports generated by GLM. The breakdown of the error types by entire articles and atomic claims is presented in Table 1. Note that some articles or claims may contain multiple types of errors.

Generator	GPT-4o		GLM	
Error type	Whole articles	Atomic claims	Whole articles	Atomic Claims
Event	38	313	36	332
participant	23	129	22	167
time	36	302	36	329
location	23	215	23	256

Table 1: Count of error types in entire articles and atomic claims

3.3 Claim verification with LLMs

Both GPT-4o and GLM models are tasked with verifying the veracity of each entire article as well as each atomic claim. To assess claim veracity, we prompted GPT-4o and GLM to evaluate the accuracy of all 182 news articles and their corresponding atomic claims. The following are the example prompts we use for the evaluation:

- **Article-level prompt:** "Today is August 1st, 2024. You are a helpful assistant that performs the below tasks: verify if the following news is accurate or false. Respond as concisely as possible."
- **Claim-level prompt:** "Today is August 1st, 2024. You are a helpful assistant that performs the below tasks: verify if the following claim extracted from a news report is accurate or false. Respond as concisely as possible."

The models are first prompted to assess the veracity of each entire article and provide a rationale for their evaluations. They are then prompted to evaluate the veracity of each atomic claim extracted from the articles, along with a rationale for each assessment. Three different prompting approaches are used in this pipeline.

3.3.1 Deterministic Prompting (Temperature 0.0)

We prompt the models to provide a singular, deterministic evaluation for each article or claim. Setting temperature to 0 minimizes randomness and allows us to observe the models' baseline claim verification performance under controlled conditions.

3.3.2 Self-consistency Prompting (Temperature 1.0)

We use a higher temperature setting (1.0) to introduce variability in the responses of the models. Models are prompted multiple times (5 times per article / claim in our experiment), and a majority voting mechanism is used to determine the final assessment. This setting simulates the potential variability in model reasoning and robustness across multiple prompts.

In each instance, the model outputs a determination (correct or false) along with a rationale for its assessment. These rationales are crucial for error analysis, offering insights into whether the model's reasoning aligns with the factual basis of the claim.

3.3.3 RAG Prompting

We queried the Google Search Serper API with manually extracted atomic claims and incorporated the retrieved results into the prompt for GPT-4 when evaluating the veracity of claims within a Retrieval-Augmented Generation (RAG) framework. The goal of this experiment was to assess whether providing search results improves the evaluation accuracy of LLMs. Due to cost constraints and the length limitation of the search engine, we did not perform this experiment with the entire article. Instead, we focused on atomic claims extracted from news reports generated by GPT-4 itself, assuming the results would generalize to other settings.

3.4 Comparing model verification with human judgments

To validate the models' evaluations, we manually verify each claim by conducting targeted web searches and cross-referencing the findings with our existing information. We use independent online sources, including reputable news databases, fact-checking websites, and government records. The human judgments serve as the gold standard for evaluating model assessments, enabling us to quantify both false positives and false negatives in the models' evaluations. Additionally, we performed error analysis to understand whether the type of news (local vs. national) and the type of claim (states vs events, true vs false claims) had a measurable impact on the model's performance. Special attention was paid to cases where the models provided no assessment, incorrect reasoning, or inaccurate evaluations.

4 Experiments

We conduct a comprehensive set of experiments to evaluate the performance of GPT-4o and GLM models in verifying claims within generated news articles. Both models are assessed in the contexts of local and national news generation, with claim verification performed across all relevant dimensions. For the claim verification task, we classify the assessment results into five possible categories, as outlined below:

- **Correct Assessment (CA):** The model correctly identifies the veracity of the claim without providing a rationale.
- **Correct Assessment and Correct Reasoning (CA/CR):** The model correctly identifies the veracity of the claim and provides a correct justification for its assessment.
- **Correct Assessment and Wrong Reasoning (CA/WR):** The model correctly classifies the claim but with flawed reasoning.
- **Wrong Assessment (WA):** The model incorrectly classifies the veracity of the claim.
- **No Assessment (NA):** The model fails to provide any assessment.

For examples of each type of assessment, please see Appendix A.5.

4.1 Entire news articles

Table 2 presents the performance data of GPT-4 (gpt-4o-20240806 and gpt-4-turbo-20240409) and GLM-4 (GLM-4-0520) in evaluating entire articles. Both models were prompted to generate news reports, followed by self-evaluation and cross-evaluation of the generated articles.

Generator	GPT-4o			GLM		
Evaluator	GPT-4o	GPT-4-turbo	GLM-4	GPT-4o	GPT-4-turbo	GLM-4
CA	1	1	0	1	0	0
CA/CR	33	28	37	33	32	32
CA/WR	9	3	4	10	7	5
WA	24	10	12	14	3	14
NA	25	50	39	32	48	39
Total	92	92	92	90	90	90

Table 2: Count of LLM-generated articles for each assessment category

Among the three models, GLM-4 achieves the highest number of correct assessments while maintaining a moderate error rate. GPT-4o produces a similar number of correct judgments but with significantly more incorrect assessments, indicating a more aggressive evaluation style.

In contrast, GPT-4 Turbo is more likely to refrain from making assessments, reflecting a more cautious approach compared to GPT-4o and GLM-4. This suggests that GPT-4-turbo prioritizes minimizing errors, even if it results in fewer overall judgments.

4.2 Individual atomic claims

In evaluating LLMs in verifying atomic claims, we conducted experiments with GPT-4o and GLM-4 to ensure our findings are generalizable across LLMs. The performance of GPT and GLM models was assessed across different temperature settings to better assess their strengths and limitations in claim verification tasks. Both models were tasked with verifying the veracity of claims extracted from LLM-generated news articles, with their evaluations measured using the identical 5-dimensional protocol we use for entire articles.

The assessment results are presented in Table 3 and we can make severval key observations. First, GPT-4o consistently provides more correct assessments (including those with and without correct reasoning) than GLM, regardless of whether it is evaluating claims from articles it generated or those generated by GLM. This trend holds across all temperature settings. Interestingly, GLM tends to produce more incorrect assessments (WA) when evaluating its own generations, while GPT-4o shows the opposite trend. The most notable finding is the high number of cases with no assessment (NA), with GLM consistently shows a slightly higher rate

Generator	GPT-4o				GLM			
Evaluator	**GPT/0**	**GPT/1**	**GLM/0**	**GLM/1**	**GPT/0**	**GPT/1**	**GLM/0**	**GLM/1**
CA (%)	60 (5.1)	72 (6.1)	6 (0.5)	9 (0.8)	15 (1.3)	14 (1.2)	8 (0.7)	12 (1.1)
CA/CR (%)	420 (35.4)	386 (32.6)	325 (27.4)	349 (29.4)	428 (37.6)	430 (37.8)	363 (31.9)	357 (31.4)
CA/WR (%)	129 (10.9)	124 (10.5)	171 (14.4)	129 (10.9)	77 (6.8)	61 (5.4)	250 (22.0)	208 (18.3)
WA (%)	106 (8.9)	134 (11.3)	63 (5.3)	68 (5.7)	128 (11.3)	104 (9.1)	35 (3.1)	49 (4.3)
NA (%)	471 (39.7)	470 (39.6)	621 (52.3)	631 (53.2)	489 (43.0)	528 (46.4)	481 (42.3)	511 (44.9)
Total	1186	1186	1186	1186	1137	1137	1137	1137

Table 3: Count and percentage of individual atomic claims for each assessment category across models at different temperature settings. GPT/0 and GPT/1 indicate GPT at temperature 0 and 1 respectively. Similarly, GLM/0 and GLM/1 indicate GLM at temperature 0 and 1.

of no assessments (NA) than GPT-4o (typically around 3–6 percentage points more) across settings. Additionally, GLM exhibits higher rates of correct assessments with wrong reasoning (CA/WR) than GPT-4o in all configurations, suggesting that its correct judgments are more often accompanied by flawed or incomplete rationales.

4.2.1 Claims in National vs Local news stories

We also attempted to evaluate the ability of LLMs to assess claims in national and local news stories. The following are example claims from national and local news stories we generated with LLMs:

- **Claims in local news:** The free rave hosted by Watertown, MA on July 15, 2024 will be held at Arsenal Park.
- **Claims in national or international news:** The 2024 Paris Olympics opening ceremony is set to take place on July 26.

A comparative error analysis of GPT and GLM models when evaluating claims from national and local news sources across different temperature settings is shown in Table 6 from Appendix A.4.1. Errors in assessments include cases where the model provides the correct assessment with wrong reasoning (CA/WR), wrong assessment (WA), or no assessment (NA). As we can see from the table, while GPT slightly outperforms GLM as indicated by the generally lower number of errors, the error rate is relatively consistent across temperatures.

The most notable finding is the substantial difference in error rates between the models' assessments of claims from national and local news, with significantly higher error rates for local news than national news. One possible explanation is that claims in national news often pertain to major events or widely recognized topics that are well-documented across diverse online sources, making these claims more likely to appear in the models' training data and thus easier to assess. In contrast, claims in local news may involve niche, region-specific issues that receive limited attention and documentation, leaving the models less prepared to verify such claims accurately. This discrepancy highlights how the scope and distribution of training data can impact the models' performance in evaluating claims with different degrees of specificity and familiarity.

4.2.2 Assessment of true claims vs false claims

We also analyze whether the LLMs make accurate or inaccurate assessments when presented with claims that are either true or false. Correct Assessment includes cases where (i) the claim is factually true, and the LLM assesses it as true. (ii) The claim is factually false, and the LLM assesses it as false. And wrong assessment includes cases where (i) the claim is factually false, but the LLM assesses it as true and (ii) the claim is factually true, but the LLM assesses it as false. We aim to investigate whether there is a difference in the accuracy with which LLMs assess factually true versus false claims. Our hypothesis is that factually true claims are more likely to be represented in the training data than factually false ones, making it more probable that factually false claims will be incorrectly assessed. Our hypothesis is born out, as results in Table 7 from Appendix A.4.2 show that both the GPT and GLM generally have a higher rate of correct assessments when the claim was factually correct while both models struggle with factually wrong claims and made wrong assessments. Among all the cases where the model made correct assessments but provided incorrect reasoning, a considerable portion of them is from claims that are factually wrong. This suggests that while the model can arrive at the correct conclusion, its internal logic or justifications may still be flawed, which happens mostly when the claims are factually incorrect.

4.2.3 State and event claims

We also experimented with asking LLMs to assess claims that are linguistic states and those that are not. Here, a state refers to a specific condition or phase in the existence of something, characterized by stability and consistency over time, whereas a non-state claim typically involves an event, signifying a significant occurrence that brings about change. A non-state claim is typically associated with a time, location, and participants. The following shows example claims categorized as state and non-state:

- **State claim:** Aryna Sabalenka is Belarusian.
- **Non-state claim:** The 2024 Australian Open Women's final was held at Margaret Court Arena on January 27.

We hypothesize that LLMs perform better on state claims because states are more stable and likely to be documented in training data, whereas events are often new and undocumented. Consequently, LLMs are more prone to errors, including wrong assessments (WA) and no assessments (NA), when evaluating non-state claims, as supported by the higher error rates observed for these claims. This hypothesis is largely born out by the higher error rate for non-states than states. We also observed a significant temperature effect and found that higher temperatures yield better results for state claims, potentially due to improved pattern recognition from broad, consistent data, while for non-state claims, the same high temperatures lead to worse outcomes as they inhibit the verification of event-specific details, causing increased uncertainty and wrong assessments. This confirms a fundamental limitation of LLMs in its knowledge of recent and dynamic events that researchers are trying to address (Ding et al., 2023; Yuan et al., 2024). More information about this can be found in Appendix A.3.

4.2.4 Fact-checking with Retrieval Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) has emerged as a popular method for fact-checking (Rothermel et al., 2024; Khaliq et al., 2024; Raina and Gales, 2024; Ullrich et al., 2024; Adjali, 2024), particularly when LLMs struggle to find information relevant to a given claim. The process typically involves transforming the claim into questions that can be used to query a knowledge source, such as the entire Internet or specific repositories like Wikipedia. The retrieved results, combined with the original claim, are then used to prompt an LLM to determine whether the claim is supported or refuted by the evidence. Additionally, the LLM can conclude that there is insufficient evidence to either support or refute the claim.

In the RAG approach, each claim is treated as a search query to retrieve relevant supporting or contradictory information from the Internet. Specifically, the claim is then fed into a Serper API to fetch relevant results from online sources. The results are then filtered to ensure relevance. For textual search results, the top k = 5 entries are selected, prioritizing those with detailed snippets, titles, and links. For knowledge graph data, attributes like titles, entity types, and descriptions are processed into usable snippets. The retrieved snippets and contextual data are consolidated and formatted into a coherent input prompt for GPT-4o. See Appendix A.2 for an example prompt.

The assessment results using the RAG approach are shown in Table 4. Compared to the non-RAG setting, the number of correct assessments (CACR) increases significantly by 21.4%, but the number of wrong assessments (WA) also rises by 16.2%, from 8.9% to 25.1%. Meanwhile, the number of no assessments (NA) drops dramatically, from 39.7% to 11.2%. At the same time, the number of correct assessments with flawed reasoning (CAWR) drops substantially from 10.9% to 1.7%, suggesting that RAG also improves the model's ability to justify its answers accurately. These results suggest that when augmented with retrieval results, GPT-4o adopts a more aggressive approach in making assessments.

Interestingly, while GPT-4o is much more likely to give a "No Assessment" when no search results (NS) are returned (37.0%), it also occasionally refrains from making a judgment (8.3%) even when retrieved search results (S) are available. This occurs when the LLM determines that the retrieved evidence is insufficiently relevant to support a definitive evaluation. Conversely, GPT-4o is capable of making correct assessments even when no relevant evidence is retrieved. A possible explanation may lie in the structure of the prompt given to the LLM. The sentence "Here are the related search snippets" followed by an empty list might implicitly signal to the LLM that no evidence supports the claim, prompting it to guess that the

Generator	GPT-4							
Evaluator	**GPT/0 Non-RAG**			**GPT/0 RAG**				
Results	Subt. (%)	FC (%)	FW (%)	Subt. (%)	FC (%)	FW (%)	S (%)	NS (%)
CA (CR)	480 (40.5)	210 (67.5)	270 (30.9)	735 (61.9)	244 (78.5)	491 (56.1)	677 (63.5)	58 (48.7)
CAWR	129 (10.9)	5 (1.6)	124 (14.2)	20 (1.7)	1 (0.3)	19 (2.2)	14 (1.3)	6 (5.0)
WA	106 (8.9)	59 (19.0)	47 (5.4)	298 (25.1)	34 (10.9)	264 (30.2)	287 (26.9)	11 (9.2)
NA	471 (39.7)	37 (11.9)	434 (49.6)	133 (11.2)	32 (10.3)	101 (11.5)	89 (8.3)	44 (37.0)
Total	1186	311	875	1186	311	875	1067	119

Table 4: Comparison between RAG and non-RAG performance with GPT4-o at Temperature 0. "S" indicates search results are returned by the Google Serper API and "NS" means no results are returned.

claim is false. However, it is debatable whether we want the LLM to make guesses this way when acting as a fact-checking system, where credibility is paramount.

5 Discussion

In our evaluation of LLMs' ability to assess the veracity of LLM-generated news articles and claims, we find that LLMs perform better when evaluating claims in national news compared to local news. They are also more accurate at assessing factually correct claims than factually wrong ones. Additionally, LLMs excel at evaluating claims expressed as linguistic states rather than those describing dynamic events. These seemingly distinct observations can be traced back to a common underlying factor: LLMs are more effective at processing well-documented, high-frequency information that is more likely to have been included in their training data. National news claims are typically better documented than local news claims, linguistic states are more stable and frequently recorded than rapidly evolving dynamic events, and factually accurate claims are more likely to appear in the training data than factually false ones.

RAG significantly improves the number of correct assessments, but also increases incorrect ones—due to irrelevant search results (56/298 cases), no retrievals (11/298), or flawed reasoning (231/298). Examples appear in Appendix A.6. The high rate of reasoning errors underscores the need to enhance LLM reasoning capabilities.

Despite RAG, many claims remain unassessed due to missing or noisy evidence. This highlights a core challenge in fact-checking news reports, which often describe novel events not covered by existing sources. Human-in-the-loop systems may be essential for verifying such content.

In cases where the assessment is correct but the reasoning is flawed (CAWR), we find that 8 out of 20 result from irrelevant retrievals, 6 from missing evidence, and 6 from pure reasoning errors—where the model misapplies logic despite having relevant information. These patterns suggest that weak or absent evidence is a major source of flawed justifications. Further analysis shows that without reliable evidence, LLMs often resort to speculative, inconsistent, or overgeneralized reasoning. When retrieval returns misleading or loosely related documents, models may incorporate incorrect details, compounding the problem. These findings highlight a key challenge for RAG systems and emphasize the need for future research to focus on improving the precision and reliability of retrieved evidence in fact-checking pipelines.

Overall, our results highlight key challenges in developing fully automatic or human-in-the-loop systems for fact-checking LLM-generated news. These include accurately extracting and decontextualizing *all and only* checkable atomic claims—a difficult task given the vague and subjective language in news stories—and improving both retrieval quality and LLM reasoning to build trust in such systems.

6 Conclusion and Future Work

We conducted a diagnostic evaluation of LLMs and RAG systems for fact-checking claims in machine-generated news. While they correctly assess nearly 65% of claims, many are misclassified or left unassessed due to irrelevant retrievals, flawed reasoning, or lack of evidence—especially for rare claims, which are common in news. Our findings highlight the need for more accurate retrieval, improved LLM reasoning, and human-in-the-loop methods for unverifiable content. Future work will focus on enhancing retrieval and reasoning capabilities, and developing hybrid systems for fact-checking LLM-generated reports.

Limitations

As a diagnostic evaluation of current challenges in fact-checking machine-generated news, our study relies on manually extracted claims—a time-consuming and labor-intensive process that limits dataset size and the scope of analysis. Despite this, we carefully curated the dataset to reflect the types of claims commonly found in LLM-generated news. In this revised version, we doubled the dataset from 92 to 182 articles and from 1,337 to 2,323 atomic claims. Our experiments, re-run on the expanded set, yielded consistent conclusions with the previous version.

Ethical Statement

Machine-generated news reports can pose significant risks if they are mistaken for authentic, factual content. To mitigate these risks, when releasing the dataset for our study, we will ensure that it is clearly labeled as machine-generated and explicitly highlight that it contains false claims. This labeling is critical to prevent misuse of the dataset and to maintain transparency for researchers, developers, and the broader community. By doing so, we aim to promote ethical research practices and minimize any potential harm arising from the dissemination of this data.

We plan to release the dataset publicly upon publication of this paper.

References

Omar Adjali. 2024. Exploring retrieval augmented generation for real-world claim verification. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 113–117, Miami, Florida, USA. Association for Computational Linguistics.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Zhenyun Deng, Michael Schlichtkrull, and Andreas Vlachos. 2024. Document-level claim extraction and decontextualisation for fact-checking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11943–11954, Bangkok, Thailand. Association for Computational Linguistics.

Zifeng Ding, Zongyue Li, Ruoxia Qi, Jingpei Wu, Bailan He, Yunpu Ma, Zhao Meng, Shuo Chen, Ruotong Liao, Zhen Han, and Volker Tresp. 2023. Forecasttkgquestions: A benchmark fornbsp;temporal question answering andnbsp;forecasting overnbsp;temporal knowledge graphs. In *The Semantic Web – ISWC 2023: 22nd International Semantic Web Conference, Athens, Greece, November 6–10, 2023, Proceedings, Part I*, page 541–560, Berlin, Heidelberg. Springer-Verlag.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9367–9385, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *ArXiv*, abs/2305.11738.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. *Preprint*, arXiv:2211.08411.

Mohammed Abdul Khaliq, Paul Yu-Chun Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. 2024. RAGAR, your falsehood radar: RAG-augmented reasoning for political fact-checking using multimodal large language models. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 280–296, Miami, Florida, USA. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Giovanni Puccetti, Anna Rogers, Chiara Alzetta, Felice Dell'Orletta, and Andrea Esuli. 2024. Ai "news" content farms are easy to make and hard to detect: A case study in italian. *Preprint*, arXiv:2406.12128.

Vatsal Raina and Mark Gales. 2024. Question-based retrieval using atomic units for enterprise RAG. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 219–233, Miami, Florida, USA. Association for Computational Linguistics.

Mark Rothermel, Tobias Braun, Marcus Rohrbach, and Anna Rohrbach. 2024. Infact: A strong baseline for automated fact-checking. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 108–112.

Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos. 2024. The automated verification of textual claims (AVeriTeC) shared task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 1–26, Miami, Florida, USA. Association for Computational Linguistics.

Herbert Ullrich, Tomáš Mlynář, and Jan Drchal. 2024. AIC CTU system at AVeriTeC: Re-framing automated fact-checking as a simple RAG task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 137–150, Miami, Florida, USA. Association for Computational Linguistics.

Yuxia Wang, Revanth Gangi Reddy, Zain Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, et al. 2024. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14199–14230.

Moy Yuan, Eric Chamoun, Rami Aly, Chenxi Whitehouse, and Andreas Vlachos. 2024. PRobELM: Plausibility ranking evaluation for language models. In *First Conference on Language Modeling*.

Moy Yuan and Andreas Vlachos. 2024. Zero-shot fact-checking with semantic triples and knowledge graphs. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 105–115, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

A.1 Example claims

1. TotalEnergies is the title sponsor for the TotalEnergies BWF Thomas & Uber Cup Finals 2024.
2. Three separate shark attacks have been reported off the coast of Maine from June 30 to July 4, 2024.
3. The United States has reported a 10% growth in Gross Domestic Product (GDP) for the fiscal year 2024 on May 29, 2024.
4. The TotalEnergies BWF Thomas & Uber Cup Finals 2024 was held at the Chongqing Olympic Sports Center.
5. The opening ceremony of the 2024 Summer Olympics was held at the Bangkok Olympic Stadium on May 29, 2024.
6. The 2024 Australian Open Women's Final was held at the Margaret Court Arena on January 27, 2024.
7. Spain won over France in the 2024 UEFA European Championship semi-final at the BVB Stadion Dortmund on July 9, 2024.
8. Zheng Qinwen is competing for her first Grand Slam final at the 2024 Australian Open Women's final at Rod Laver Arena on January 20, 2024.
9. The free rave hosted by Watertown, MA on July 15, 2024 will be held at Arsenal Park.
10. The discovery of the new COVID-19 in New Hampshire variant was announced by health officials on May 29, 2024.
11. The Dallas Mavericks won Game 4 of the 2024 NBA Finals in overtime at the American Airlines Center in Dallas, Texas.
12. The 2024 NBA Finals Game 7 was played at the American Airlines Center in Dallas on May 29, 2024.
13. Kobe Bryant has announced his return to the Los Angeles Lakers on May 29, 2024.
14. On July 3, 2024, FIFA has announced that London, United Kingdom, will host the 2026 FIFA World Cup.
15. On May 29, 2024, Stephen Curry was traded from the Golden State Warriors to the Chicago Bulls.

A.2 Example RAG prompt

The prompt structure includes: The original claim. The concatenated evidence snippets from the retrieved results. And a preamble describing the task (e.g., assessing the factual accuracy of the claim). Example Prompt: "The following claim needs to be evaluated for accuracy: 'CLAIM'." "Here are the related search snippets: EVIDENCE-TEXT." "Based on the snippets provided, evaluate whether the claim is accurate or false. " "Provide a clear and reasoned explanation."

A.3 Assessments for State and Event Claims

Table 5 presents a comparison between the evaluation performance of LLMs on state and non-state (event) claims. LLMs are better at assessing state claims than non-state claims, as indicated by the generally lower number of WA and NA cases for state claims and higher number of such cases for non-state claims.

There is also a significant temperature effect. For state claims, which often pertain to more standardized and systemic issues, higher temperatures might enhance the model's ability to identify patterns and make accurate assessments. These claims are typically based on broader, more consistent data that may not be as sensitive to small fluctuations or variability in the input data. Conversely, higher temperatures introduce greater variability in responses, which impacts non-state claims differently.

Generator	GPT-4o				GLM			
Evaluator	GPT/1	GPT/0	GLM/1	GLM/0	GPT/1	GPT/0	GLM/1	GLM/0
State								
CA (%)	36 (11.5)	22 (7.0)	4 (1.3)	5 (1.6)	14 (6.6)	6 (2.8)	7 (3.3)	5 (2.3)
CA/CR (%)	173 (55.3)	144 (46.0)	177 (56.5)	115 (36.7)	152 (71.4)	124 (58.2)	124 (58.2)	93 (43.7)
CA/WR (%)	5 (1.6)	5 (1.6)	5 (1.6)	13 (4.2)	1 (0.5)	16 (7.5)	3 (1.4)	10 (4.7)
WA (%)	54 (17.3)	46 (14.7)	45 (14.4)	42 (13.4)	10 (4.7)	13 (6.1)	25 (11.7)	19 (8.9)
NA (%)	45 (14.4)	96 (30.7)	82 (26.2)	138 (44.1)	36 (16.9)	54 (25.4)	54 (25.4)	86 (40.4)
Subtotal	313	313	313	313	213	213	213	213
Non-State								
CA (%)	36 (4.1)	38 (4.4)	5 (0.6)	1 (0.1)	0 (0.0)	9 (1.0)	5 (0.5)	3 (0.3)
CA/CR (%)	213 (24.4)	276 (31.6)	172 (19.7)	210 (24.0)	278 (30.1)	304 (32.9)	233 (25.2)	270 (29.2)
CA/WR (%)	119 (13.6)	124 (14.2)	124 (14.2)	158 (18.1)	60 (6.5)	61 (6.6)	205 (22.2)	240 (26.0)
WA (%)	80 (9.2)	60 (6.9)	23 (2.6)	21 (2.4)	94 (10.2)	115 (12.4)	24 (2.6)	16 (1.7)
NA (%)	425 (48.7)	375 (42.9)	549 (62.9)	483 (55.3)	492 (53.2)	435 (47.1)	457 (49.5)	395 (42.8)
Subtotal	873	873	873	873	924	924	924	924
Total	1186	1186	1186	1186	1137	1137	1137	1137

Table 5: A comparison of state vs non-state claims.

Non-state claims defined by more dynamic, event-specific details like timing, location, or participants, are harder for the models to verify with confidence under higher temperatures. The randomness at this setting leads to the model producing a broader array of responses, which is beneficial for creativity but not ideal for precision. In fact, the variability might cause the model to contradict itself or lose consistency, particularly when precise details are required to confirm an event. This can explain the higher no assessments for non-state claims under high-temperature settings, as the models struggle with conflicting or incomplete information about specific events. Interestingly, while lower temperatures generally lead to more correct assessments (CA and CA/CR), they also tend to increase the number of correct assessments with wrong reasoning (CA/WR)in non-state claims. This suggests that although models may make more definitive judgments at lower temperatures for event-specific information, the quality or consistency of the supporting rationale may suffer. In contrast, high temperatures tend to reduce the ratio of flawed justifications among correct answers, even though they also bring more uncertainty. This pattern does not appear in state claims, where CA/WR is infrequent.

A.4 Supplemental tables

A.4.1 Claims in National vs Local news stories

We provide a detailed error breakdown for national and local news claim evaluation in Table 6.

A.4.2 true claims vs false claims

Table 7 presents results comparing LLMs' assessment accuracy on factually true vs false claims.

A.5 Examples of each type of assessment

- **CA:**
  - **claim:** "Kobe Bryant retired in 2016 and tragically passed away in a helicopter crash in 2020."
  - **eval:** "True."
- **CA/CR:**
  - **claim:** "Kamala Harris is U.S. President in 2024."
  - **eval:** "As of the last update, Kamala Harris is not the U.S. President; she is the Vice President. Joe Biden is the President of the United States. However, please check the most recent and reliable news sources to confirm the current office holders, as situations can change."
- **CA/WR:**
  - **claim:** "Osaka said in her post-match interview of 2024 Australian Open Women's Final, "This was one of the toughest matches I've ever played. Iga is an incredible player, and she pushed me to my limits today. Winning here at the Australian Open again is a dream come true, and I'm so grateful for the support from the fans.""

Generator	GPT-4o					GLM				
Evaluator	Subt.	GPT/0	GPT/1	GLM/0	GLM/1	Subt.	GPT/0	GPT/1	GLM/0	GLM/1
National(%)	785	374 (47.7)	407 (51.8)	496 (63.2)	475 (60.5)	756	373 (49.3)	372 (49.2)	443 (58.6)	447 (59.1)
Local(%)	401	332 (82.8)	321 (80.0)	359 (89.5)	353 (88.0)	381	321 (84.3)	321 (84.3)	323 (84.8)	321 (84.3)
Total	1186	706 (59.5)	728 (61.4)	855 (72.1)	828 (69.8)	1137	694 (61.0)	693 (60.9)	766 (67.4)	768 (67.6)

Table 6: Errors from evaluating claims in national or local news. Each cell represents the percentage of claims that are incorrectly assessed for that category (national vs local), with the last row representing the number of errors / the total claims for that evaluator.

Generator	GPT-4				GLM			
Evaluator	GPT/0		GLM/0		GPT/0		GLM/0	
Veracity	FC(%)	FW(%)	FC(%)	FW(%)	FC(%)	FW(%)	FC(%)	FW(%)
CA (CR)	210 (67.5)	270 (30.9)	168 (54.0)	163 (18.6)	152 (78.8)	291 (30.8)	111 (57.5)	260 (27.5)
CAWR	5 (1.6)	124 (14.2)	15 (4.8)	156 (17.8)	3 (1.6)	74 (7.8)	3 (1.6)	247 (26.2)
WA	59 (19.0)	47 (5.4)	51 (16.4)	12 (1.4)	9 (4.7)	119 (12.6)	24 (12.4)	11 (1.2)
NA	37 (11.9)	434 (49.6)	77 (24.8)	544 (62.2)	29 (15.0)	469 (49.7)	55 (28.5)	426 (45.1)
Total	311	875	311	875	193	944	193	944

Table 7: Comparison of LLM assessment accuracy for factually correct (FC) and factually incorrect (FW) claims with GPT and GLM as evaluators at 0 temperature.

- **eval:** “The information provided seems incorrect. The Australian Open is typically held in January, so if today is July 29th, 2024, Osaka would not be referring to a recent Australian Open victory in her interview.”

• **WA:**

- **claim:** "The XVIII International Chopin Piano Competition held in Warsaw, Poland, on October 25th, 2021 saw young pianists from around the world converge on the Polish capital to compete for the coveted first prize."
- **eval:** “The information you provided is accurate. The XVIII International Chopin Piano Competition was indeed held in Warsaw, Poland, starting on October 25th, 2021, and it attracted young pianists from around the world to compete for the prestigious first prize.”

• **NA:**

- **claim:** "The rave event called ’Watertown Rave: Lights & Beats’ on July 15, 2024 host by Watertown, MA is expected to draw a large crowd of electronic dance music (EDM) enthusiasts from across New England."
- **eval:** “As an AI, I cannot provide real-time or future event information. To verify the accuracy of news about events like the "Watertown Rave: Lights & Beats," you should check with local event listings, news sources, or the official event organizers. It’s always best to get the latest updates from reliable sources.”

A.6 Error cases in the RAG setting

• **no search results:**

- **claim:** "Watertown Town Manager Michael Driscoll said in June 2024 "This is an important development for our town, and we value the input of our residents. The public meeting will be an opportunity to hear from the university and provide feedback to ensure the project benefits both Brandeis and the Watertown community.""
- **eval:** “The claim is false. The statement was actually made by Watertown Town Council President Mark Sideris, not Town Manager Michael Driscoll.”
- **claim:** "Russell Westbrook wrote on his social media "Congrats on an amazing career, PG! Proud of everything you’ve accomplished and excited for what’s next." on May 29, 2024"
- **eval:** “The claim is false. The provided snippets indicate that Russell Westbrook’s social media post congratulat-

ing “PG” was actually directed at Paul George, who announced his retirement from professional basketball. The date of the post is also consistent with the timing of Paul George’s retirement announcement, which was reported on May 29, 2024. Therefore, the claim that Westbrook’s message was about Paul George’s career and future is accurate.”

- **irrelevant search results:**
  - **claim:** "Bruce Liu’s final round performance of the XVIII International Chopin Piano Competition held in Warsaw on October 25, 2021 featured a breathtaking rendition of Chopin’s Piano Concerto No. 1 in E minor, Op. 11."
  - **eval:** “The claim is false. The search snippets indicate that Bruce Liu performed Chopin’s Piano Concerto No. 2 in F minor, Op. 21 during the final round of the XVIII International Chopin Piano Competition, not Concerto No. 1 in E minor, Op. 11 as stated in the claim. The snippet specifically mentioning "Bruce Liu - F.Chopin Piano Concerto No.2 in F minor, Op.21" during the competition confirms this.”
  - **claim:** "The first shark attack took place on Sunday, June 30, 2024, at Old Orchard Beach, where a 35-year-old man suffered serious injuries to his leg while swimming."
  - **eval:** “The claim that the first shark attack took place on Sunday, June 30, 2024, at Old Orchard Beach is false. The search snippets indicate that there was a fatal shark attack before this date, involving Julie Dimperio Holowach, who died as a result of a shark attack while swimming in Casco Bay near Bailey Island. This incident is noted as Maine’s first-ever fatal shark attack. Therefore, the claim about the first shark attack occurring in 2024 is inaccurate, as there was at least one documented shark attack prior to this date.”
- **reasoning failure:**
  - **claim:** "Aryna Sabalenka is the second seed of the 2024 Australian Open Women’s."
  - **eval:** “The claim that Aryna Sabalenka is the second seed of the 2024 Australian Open Women’s is false. The search snippets indicate that Aryna Sabalenka was ranked No. 2 during the tournament, as mentioned in the snippet: "In Saturday’s women’s final, No. 2-ranked Aryna Sabalenka defeated Zheng Qinwen in straight sets, for her second overall major title." Being ranked No. 2 and being the second seed are typically synonymous in major tennis tournaments, as seeding is generally based on world rankings just before the event. Therefore, the claim is accurate.”