Real-time Fake News from Adversarial Feedback

Anonymous ACL submission

Abstract

We show that existing evaluations for fake news detection based on conventional sources, such as claims on fact-checking websites, result in high accuracies over time for LLM-based detectors—even after their knowledge cutoffs. This suggests that recent popular fake news from such sources can be easily detected due to pre-training and retrieval corpus contamination or increasingly salient shallow patterns. Instead, we argue that a proper fake news detection dataset should test a model's ability to reason factually about the current world by retrieving and reading related evidence. To this end, we develop a novel pipeline that leverages natural language feedback from a RAG-based detector to iteratively modify real-time news into deceptive fake news that challenges LLMs. Our iterative rewrite decreases the binary classification ROC-AUC by an absolute 17.5 percent for a strong RAG-based GPT-4o detector. Our experiments reveal the important role of RAG in both detecting and generating fake news, as retrieval-free LLM detectors are vulnerable to unseen events and adversarial attacks, while feedback from RAG detection helps discover more deceitful patterns in fake news.

1 Introduction

The spread of fake news can have serious consequences, such as influencing elections, inciting violence, and misleading critical decision-making, especially in health. In the NLP community, researchers have long focused on developing methods and evaluation for automatic fake news detection (Zellers et al., 2019). The rise of LLM has significantly changed the landscape of this problem (Goldstein et al., 2023; Chen and Shu, 2023).

LLMs have demonstrated an impressive knowledge and reasoning ability, enabling them to detect fake news with high accuracy (Chen and Shu, 2024; Pelrine et al., 2023). The pretraining and prompting paradigm of LLM reduces the risk of creating dataset-specific models that are prone to in-distribution shortcut learning (Pagnoni et al., 2022). However, due to the opaque nature of LLM training, their evaluation is often influenced by potential contamination issues, which can result in misleading outcomes and a need for out-of-distribution evaluation (Zhou et al., 2023; Vu et al., 2024; Huang et al., 2024). Existing fake news detection datasets are commonly sourced from past claims from fact-checking websites (Wang, 2017; Shu et al., 2018). These websites, such as PolitiFact and Snopes, often feature popular claims that are widely circulated on the internet, leading to potential contamination during the large-scale pre-training of LLMs.

To evaluate the LLM's ability to detect misinformation under natural temporal distribution shifts and to avoid potential contamination issues, we study *real-time fake news that occurs concurrently or after model training*. LLM identifies the plausibility of unseen misinformation by finding supporting evidence from two sources of knowledge: its internal parametric knowledge and an external knowledge through retrieval augmentation of real-time information. Using new fact-checking results from the aforementioned sources avoids contamination of the parametric knowledge but can, however, suffer from label leakage in the retrieval context. Moreover, fact-checking publishers are often biased in the selection of news they check, which can limit the evaluation to a certain type of misinformation. We find evidence of all these issues in experiments with two popular fact-checking sources. On Snopes, we observe near-perfect retrieval-free detection for past fake news and a clear decline after knowledge cutoffs, which is easily restored by retrieval augmentation. On PolitiFact, we find a surprising uptrend in retrieval-free detection performance even after the knowledge cutoff dates of models, suggesting a gain from *non-factual* salient patterns of the recent political claims selected. To avoid these issues and evaluate real-time fake news

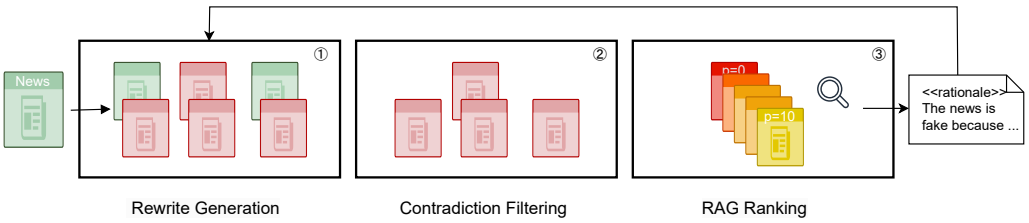

Figure 1: Our proposed adversarial iterative fake news generation. 1) A true news is rewritten by a LLM to multiple candidates containing misinformation. 2) Candidates that do not contradict the original true news are filtered out. 3) The remaining candidates are ranked on their plausibility score according to a RAG-based detector and the most plausible one is selected, the detector's rationales are used to inform the next round generation. The process can be repeated for multiple iterations.

detection with fresh information (Yang et al., 2022; Hu et al., 2023b; Liao et al., 2023), we investigate LLM-synthesized fake news.

LLM-generated fake news has shown to be more deceitful than human-written fake news (Chen and Shu, 2024); however, we find that current neural fake news still cannot fool LLM detectors. Therefore, we propose an adversarial iterative approach to generate fake news that can deceive strong detectors, inspired by LLMs' ability of continual improvement based on feedback (Madaan et al., 2023). The generator leverages feedback from an adversary retrieval-augmented detector in the form of a rationale discussing the factuality of the generated fake news, mimicking the real-world scenario in which fact-checkers provide explanations for their verdicts. Given this verdict and the fake news generated in the previous round, the generator then adds another round of modifications to the fake news such that it would fool the prior verdict. This process is repeated a few times, allowing the generator to learn from the detector's perspective. Empirically, highly performant LLM-based detectors on conventional political fact-checking struggle with the fake news generated by our approach.

Our work makes the following contributions. First, we conduct a background study of LLM detectors on PolitiFact and Snopes data over the years, revealing issues that affect the continued applicability of these popular sources. Second, we propose an iterative adversarial approach that introduces highly deceptive misinformation to real-time news from various domains. We show that new datasets created by this method pose a severer challenge than previous neural fake news datasets. The increasing deception generalizes well across different setups varying LLM backbones and retrieval contexts. Lastly, we analyze LLM behaviors regarding RAG-targeted misinformation of the current world.[1]

2 Background

2.1 Problem Formulation

We formulate the fake news detection task as a binary classification problem, where the primary objective is to determine whether a given news is genuine or fabricated. Instances of news cover a wide variety of topics, such as political claims, social media posts, and news releases. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ represent the dataset, where each x_i is a news and $y_i \in \{0, 1\}$ is a binary label indicating the factuality of the news. A detector $f : \mathcal{X} \rightarrow [0, 1]$ maps each news x_i to a probabilistic score $\hat{y}_i = f(x_i)$, reflecting the likelihood of the news being factually correct. Under our real-time setup, we also refer to this likelihood as "plausibility" as the wording allows flexibility in fact-checking current and future events where direct evidence may be unavailable. We adopt AUC-ROC as the main metric to evaluate the effectiveness of a detector as it quantifies the model's capability to classify across various thresholds.

2.2 Analysis of Fact-checking Sources

Fact-checking websites are widely used sources for obtaining news content and human labels when curating fake news detection datasets. Despite their popularity, it is unknown whether they pose continued challenges to state-of-the-art LLM fake news detectors. We study two representative sources of this kind: PolitiFact and Snopes,[2] both of which are used in creating popular fake news detection datasets (Wang, 2017; Shu et al., 2018; Shahi and

[1] We will release our code and data to facilitate further research on fake news detection and generation.

[2] politifact.com, snopes.com

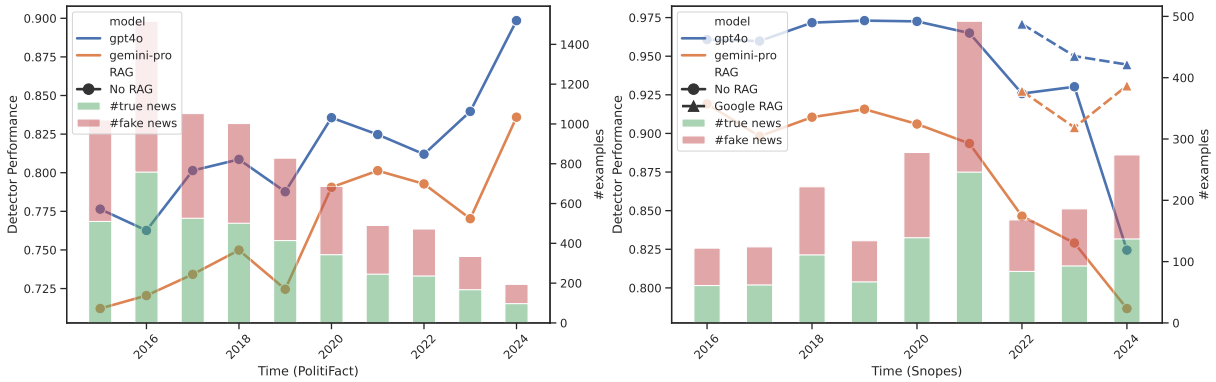

Figure 2: Comparison of different retrieval-free detectors (AUC-ROC) on PolitiFact and Snopes data over the years. The data is balanced by downsampling the majority class. Both the GPT-4o and Gemini Pro models claim to have knowledge up to the end of 2023. Recent fake news on PolitiFact is increasingly easy to detect by LLMs without the need for fresh external knowledge. While Snopes challenges LLMs in detecting up-to-date fake news, simple retrieval augmentation largely brings back near-perfect performance.

Nandini, 2020). In order to evaluate the performance of LLM-based detectors on both past and recent year data, we collect 9,244 PolitiFact news from June 2015 to August 2024, and 3,212 Snopes news from Jan 2016 to October 2024.

We run a set of LLM-based detectors using simple zero-shot prompting (detailed setup described in §4). In Figure 2, we observe two opposite trends of detector accuracies through the years. On Snopes, GPT-4o exhibits near-perfect performance on past data, showing potential data contamination due to internet-scale pre-training. Recent fake news, both near and beyond the knowledge cutoff dates, becomes more challenging to detect due to the lack of current world knowledge. However, simply augmenting the detectors with Google search results is sufficient to restore performance. On PolitiFact, a source containing mostly political claims, the performance of retrieval-free detectors improves over time. In the most recent year, 2024, we see that LLMs continue to improve, even after the knowledge cutoff and the model release dates. Pelrine et al. (2023) report similar findings. This suggests that up-to-date knowledge is not the deciding factor for PolitiFact fake news detection performance.

We further investigate the emergent patterns on PolitiFact that have increased the separability of fake and real news (detailed setups in A.1). We first confirm the trend on a much smaller LM, RoBERTa (Liu et al., 2019), which has far less world knowledge. Fine-tuned RoBERTa also performs significantly better on recent year data. We then ablate key features such as removing the originator and publish time of the news from detector inputs, and paraphrasing the claim content. None of these features affects the trend of increasing performance over the years (Figure 8). The prediction distribution in Figure 3 reveals that the changes are mostly in the detector's perception of the fake news class, which shifts towards lower scores. Manually going through some fake news examples, we find that in earlier years, political fake news contains statements that, while hyperbolic, can be anchored in a context that requires research through reliable sources (e.g., public records or legislative history) to verify. In contrast, recent fake news includes more sensational and less easily verified claims that seem more speculative or exaggerated without immediate substantive evidence. Consequently, the detection of recent PolitiFact fake news requires less factual knowledge and reasoning, but more pattern recognition and common sense.

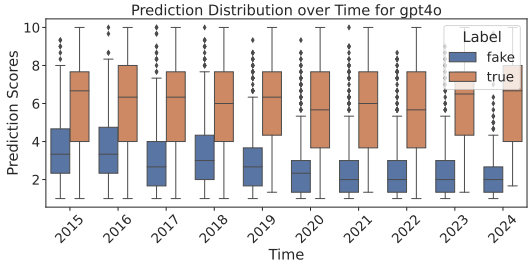

Figure 3: GPT-4o (retrieval-free) predicted plausibility of PolitiFact claims over the years. The distribution of the true news class remains relatively stable, while the distribution of the fake news class shifts significantly towards lower scores.

To summarize, evaluation on fact-checking data is susceptible to the uncontrolled biases in the cu-

ration process. Traditional evaluation of PolitiFact focused on *surface-level linguistic patterns* (Wang, 2017) is no longer challenging for strong LLMs due to increasing separability of fake political news. Other fact-checking sources (e.g., Snopes) suffer from potential contamination in both the pre-training and retrieval-augmentation stages because of the popularity of both the claims and fact-checking results on the internet. Therefore, to evaluate the factual reasoning ability of LLMs in detecting contemporary fake news, we need to explore new ways to create datasets that cover real-time information from diverse domains for a more challenging evaluation.

3 Methodology

The wide accessibility of LLMs has not only enabled strong detection models but also facilitated the mass generation of more credible and persuasive fake news (Kreps et al., 2022; Goldstein et al., 2023). Therefore, in combating these emerging threats, the evaluation of detection models should also evolve to incorporate challenging machine-generated misinformation.

One key challenge in creating fake news datasets is obtaining automatic labels for the data. Open-ended generation, although effective in creating diverse fake news, simultaneously introduces false positives that are hard to verify due to the professional skills required for fact-checking. Our proposed approach leverages the strengths of LLMs in controlled misinformation generation (Zellers et al., 2019; Chen and Shu, 2024) to introduce factual errors to real news, while maintaining its overall context and style. We implement filtering protocols as an additional safeguard to reduce invalid generation.

Figure 1 and Algorithm 1 illustrate our pipeline. Formally, we start with a trusted news corpus $\mathcal{T}$, which contains real-time news from various domains. We independently rewrite each news $\tau_i \in \mathcal{T}$ to generate multiple fake news candidates $\mathcal{F}_i$. To ensure that $\mathcal{F}_i$ contradict the original true news τ_i, we employ a LLM-based contradiction detector. We put an additional upper limit on the amount of edits allowed in the rewriting process using a threshold on the Levenshtein distance $\Delta(\tau_i, f_{ij})$, where $f_{ij} \in \mathcal{F}_i$, between the original true news τ_i and the rewritten news f_{ij}.

After filtering out the candidates that do not meet these criteria, we rank the remaining candidates

Algorithm 1 Adversarial Iterative News Rewriting

Require: $TrueNews$, k {k is the maximum number of iterations}
$CurrentFake \leftarrow TrueNews$
for $i \leftarrow 1$ **to** k **do**
  $Candidates \leftarrow$ GenerateCandidates($CurrentFake$)
  repeat
    $Filtered \leftarrow \emptyset$
    for each c_i **in** $Candidates$ **do**
      if ContradictsOriginal(c_i, $TrueNews$) **and** IsWithinEditLimit(c_i, $TrueNews$) **then**
        $Filtered \leftarrow Filtered \cup \{c_i\}$
      end if
    end for
    if $Filtered = \emptyset$ **then**
      $Candidates \leftarrow$ GenerateCandidates($CurrentFake$)
    end if
  until $Filtered \neq \emptyset$
  $Ranked \leftarrow$ RankByPlausibility($Filtered$)
  $CurrentFake \leftarrow$ SelectMostPlausible($Ranked$)
end for
return $CurrentFake$

based on the plausibility score provided by a fake news detector $g(f_{ij}|c)$, where c is the optional external context retrieved by a retrieval model $\mathcal{R}(f_{ij})$. If none of the candidates meet the criteria, the generator resamples a new batch of candidates. From this ranked list, we select the top-ranked candidate as the most plausible fake news.

$$\hat{f}_i = \arg\max_{f_{ij} \in \mathcal{F}_i} g(f_{ij}|c) \quad (1)$$

$\hat{f}_i$ and the detector's rationale are then serve as additional information to inform the next round of generation on τ_i, creating an iterative process that gradually deceives the detector. In the end of the process, we obtain the most deceptive fake news $\hat{f}_i$ across all iterations.

A critical component of our approach is the retrieval-augmented detector. Using a retrieval-free detector as an adversary, the generator can only improve the factual consistency within the news content or exploit the weaknesses of a LLM with outdated internal knowledge, which can generally be seen as a case of self-evaluation (Kadavath et al., 2022). However, with the information from the retrieved external context, the generator can learn to deceive cross-verification and improve factual consistency beyond its own knowledge, which is important for our real-time news rewrite setup.

One implementation challenge in Eq. 1 is that we need to run retrieval for each candidate to obtain the external context. This can be computationally expensive for sophisticated retrievers and costly if commercial APIs are used. In practice, we retrieve

the external context only for the highest-ranked candidates at the end of each iteration, which we use to inform the next round of detection. Because of the constraints on the amount of change in each iteration, the external context is expected to remain relevant for the next batch generation. In final evaluation, we rerun the retrieval each time to ensure the external context is up-to-date.

3.1 Dataset Creation

We first obtain 431 actual news stories from NBC News using the *news-please* crawler (Hamborg et al., 2017). These news stories are from March 1 to March 13, 2024, covering domains such as politics, business, sports, U.S., and world news. Compared to claims from fact-checking websites, these news stories are roughly twice the length of the content. They can be seen as long-form claims, which can be more difficult to fact-check. This two-week range is close to the time when we start the experiments and is beyond the knowledge cutoffs of the LLMs we use, thus ensuring no contamination in model training. Our code supports replicating this process for other date ranges.

This set of news stories serves as seed true news for our generation pipeline and undergoes one-to-one rewriting to obtain 431 fake news instances. We manually examine all real-fake pairs of the final-round rewrite and filter out 29 invalid pairs. These generations manage to bypass our contradiction detector without introducing factual errors that contradict claims in the corresponding real news. Comparing 100 pairs from the first and last rounds of rewrites, the ratio of failed rewrites remains stable at around 7%. We also fact-check 100 last-round examples under an unpaired, shuffled, label-hidden setup, establishing 99% human performance with access to Google. Our final dataset consists of 402 true news and 402 fake news.

4 Experimental Setup

4.1 Generator Setup

We adopt GPT-4o with different prompts for the all three roles (i.e., generator, contradiction detector, and reranker) in the generation pipeline. The pipeline is iterated for 6 rounds with an extra preparation round 0 of a direct rewrite (no rationale and ranking) on the seed true news. We instruct GPT-4o to "*introduce some believable factual errors*" without mentioning any concrete strategies or constraints, leaving the candidate generation process open-ended. The generator produces 8 candidates in a zero-shot chain-of-thought fashion for each news story in each round, which are then filtered and ranked based on the plausibility score from the detector. The contradiction detector produces 10 binary scores for each candidate, only if more than eight of them are positive, the candidate is considered to contradict the original true news. String edit distance is used to limit the amount of change in each iteration, with a threshold of more than 60% overlapped tokens. A GPT-4o-based detector (detailed in the next section) is used to provide rationales and serves as the reranker to select the best generation after contradiction filtering.

4.2 Detector Setup

For implementing the LLM detectors, we simply instruct the models to produce plausibility scores (1-10 from implausible to plausible) for the given news in a 0-shot setting. The score is mapped to [0, 1] when computing metrics. We use the directive "*Today is March 26, 2024. You predict the plausibility of a news you haven't seen*" to inform the model about the current date and the unseen nature of the news story. Using "plausibility" instead of "factuality" is empirically more effective for recent and future events, vice versus for past events (Table 9). We sample multiple plausibility scores from the unscaled LLM prediction distribution (temperature $t = 1$) and then take an average.

We employ two retrievers for RAG-based detection: an in-house *News* corpus consisting of 811k news articles from various news sources within March 2024, indexed by a pre-trained DPR (Karpukhin et al., 2020) retriever; real-time *Google* search results provided by SerpApi. We extensively filter both sources to remove the exact news and all results from NBC News. From either source, five relevant instances can be obtained using the news headline as a search query. The results are then inserted into the prompt for fact-checking. *News* is used for the GPT-4o detector in the generation pipeline. *Google* is only used in evaluation due to the cost of the API.

Our GPT-4o detector is comparable to recent state-of-the-art RAG detectors on the popular LIAR datasets (Table 2). Final evaluation is conducted on multiple LLM detectors, including GPT-4o, GPT-3.5, Gemini Pro and Flash (Gemini Team, 2024), and open-source 405B Llama 3.1 (Meta AI, 2024). Detailed versions and knowledge cutoffs are provided in Table 6. The actual prompts used by all

Retriever	None			News			Google		
Detector	first	last	Δ	first	last	Δ	first	last	Δ
Gemini-Flash	57.0	50.7	6.3	76.1	62.4	13.8	84.1	76.6	7.5
Gemini-Pro	58.6	51.6	7.0	74.9	62.8	12.1	82.6	73.7	8.9
GPT-3.5	53.7	50.3	3.4	69.3	57.6	11.7	78.2	69.3	8.9
GPT-4o	58.5	48.8	9.7	82.4	64.9	17.5	93.1	86.1	7.0
Llama 3.1	60.5	54.0	6.6	81.3	67.4	13.9	93.3	86.6	6.7

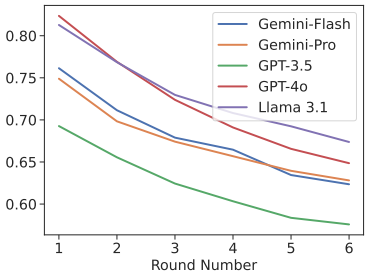

Table 1: AUC-ROC scores of different detectors on the generated fake news of the first and last iteration. The Δ column shows the effects of the iterative process. *News* refers to the in-house DPR retriever on news-please data, and *Google* refers to the Google search API results. The right figure presents results across rounds under the *News* retrieval setting–more iterations lead to stronger deception.

Method	LIAR		
	AUC	F1-Ma	Acc
RoBERTa-L (2023)	-	64.7	64.1
GPT-4 (2023)	-	68.1	68.2
MUSER (2023)	-	64.5	-
STEEL (2024)	-	71.4	-
GPT-4o	77.5	70.7	70.9
GPT-4o + Google	81.1	74.7	75.6

Table 2: Our detectors demonstrate superior performance to the state of the art on the LIAR dataset (Wang, 2017). MUSER and STEEL have access to multi-step retrieval on Wikipedia and Bing, respectively.

	2023b		2024		Ours	
	G++	P++	Rewrite	Mani.	First	Last
Gemini-Flash	72.7	82.8	80.4	75.3	57.0	50.7
Gemini-Pro	72.8	86.2	76.6	77.3	58.6	51.6
GPT-3.5	64.5	77.4	81.2	68.9	53.7	50.3
GPT-4o	81.8	88.8	84.3	85.4	58.5	48.8
Llama 3.1	80.4	91.3	84.0	83.3	60.5	54.0

Table 3: AUC-ROC scores of different retrieval-free detectors on recent LLM-generated fake news datasets (Su et al., 2023b; Chen and Shu, 2024). Our pipeline produces significantly more difficult datasets.

components are provided in Appendix A.4.

4.3 Baseline Datasets

We compare to two recent LLM fake news datasets. Su et al. (2023b) apply open-ended rewriting to fake news and rephrase real news from GossipCop and PolitiFact. Chen and Shu (2024) explore various misinformation generation approaches with LLMs, including rewriting fake news and targeted information manipulation of true news. Both approaches utilize one round of generation. Appendix A.3 provides detailed dataset statistics.

5 Experimental Results

Evaluation results (Table 1) show that our adversarial iterative generation pipeline can produce fake news that can deceive strong LLM-based detectors. We find Llama 3.1 to be the best at detecting fake news generated by our pipeline using GPT-4o. As the earliest released model in our selection, GPT-3.5 is the most vulnerable, consistent with results of well-established public benchmarks (Chiang et al., 2024). Datasets generated from our pipeline are significantly more difficult than previous neural fake news datasets (Table 3).

Through each iteration, the pipeline progressively enhances the deceptive quality of the generated fake news. Relying on the feedback from the GPT-4o RAG-based detector with the in-house retrieval corpus, the generator proves most effective at deceiving this particular detection setting, achieving a reduction of 17.5 AUC-ROC points. Nevertheless, these enhancements are shown to consistently generalize across different LLM backbones and retrieval contexts.

5.1 Analysis

Real-time news can be implausible to off-the-shelf LLM detectors. We find that true news from NBC News is generally plausible to the LLM detectors, with an average plausibility score of 7.8. However, some news stories are rated as low as 3.3, indicating that the LLM detectors can be fooled by real-time news that is beyond their knowledge (Figure 4). For example, GPT-4 finds news stories reporting on RFK Jr.'s presidential campaign and the White House's reply to the Sesame Street character's Twitter account on inflation issues to be implausible.[3]

[3]Links to the RFK news and the White House tweet.

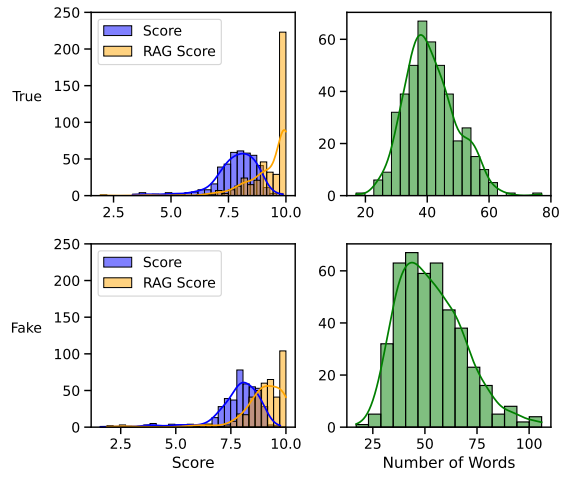

Figure 4: Distribution of the plausibility scores (retrieval-free and RAG) and the number of words in the set of real news from NBC News (positives, top) and the last-round rewritten fake news (negatives, bottom).

Retrieval free detectors are vulnerable to adversarial attacks. It is very easy to rewrite unseen news to generate fake news that can evade detection by non-RAG-based LLM detectors. This is evident from the detector performance reported in Table 1 and also from the prediction distribution (purple bars in Figure 4). Retrieval-free detectors result in near random AUC even for the first round generation. On the contrary, RAG-based detectors are generally more robust as they leverage up-to-date external knowledge. As we can clearly see from the true news prediction distribution, more than half of examples receive an unbeatable maximum plausibility score from the RAG-based GPT-4o detectors, while the retrieval-free detector is more conservative in its predictions. The average plausibility score of the RAG-based detectors on the true news is 9.3, significantly higher than 7.8 of the retrieval-free detectors. The confidence of RAG-based detectors does not come without a cost, as they also assign generally higher scores to fake news, suggesting weaknesses in the LLMs' ability to reason about future events in the external context (Shi et al., 2023).

Stronger defenders enable stronger attackers. RAG-based detectors mount an effective defense against LLM-generated fake news, but the generator can adapt to evade their detection, especially when the detector's rationale is accessible. In Table 5, We show that the RAG-based detector rationale surpasses the non-RAG rationale and ranking-only variants in the first round and reinforces its advantage through iterations. In the 6th round, the RAG-based rationale (64.9) reduces the AUC by 6.5 more than the non-RAG rationale (71.4). We run the other variants for up to 2 rounds due to resource constraints.

1. *Trump says he will debate Biden anytime*
After skipping On March 6, 2024, after participating in a few select GOP primary debates, Trump challenged Biden to step onto a stage with him before the November election. , suggesting that the debate be held at a neutral venue yet to be decided.

2. *Violence, hunger Rising violence, growing food shortages, and unstable missing political leadership*
A long-simmering severe crisis in Haiti has come to a head reached alarming levels as its leader remains stranded in Puerto Rico while its people starve the Dominican Republic for several days. Citizens face extreme food shortages and live in fear of rampant escalating gang-related violence.

3. (round 1) *Iranian parliament Saudi Arabia's parliamentary vote sees a low turnout despite government push*
Iran Saudi Arabia held its first parliamentary election Friday since mass 2022 protests over mandatory hijab laws following the death of Mahsa Amini, apparently drawing a low turnout amid calls for a boycott.
RAG Rationale Saudi Arabia does not have a parliamentary system; it is an absolute monarchy with a Consultative Assembly (Shura Council) that is appointed by the King, not elected by the public. The reference to Mahsa Amini and mandatory hijab laws is more relevant to Iran, not Saudi Arabia, indicating a possible mix-up of countries. There is no mention of a Saudi parliamentary election in the related news stories, which would be a significant event likely covered by major news outlets.
RAG Score 2.06

3. (round 3) *Iranian parliament parliamentary vote sees a low turnout despite government push*
Iran held its first parliamentary election Friday since mass 2022 protests over mandatory hijab laws fuel price hikes following the death of Mahsa Amini, government's announcement, apparently drawing a low turnout amid calls for a boycott.
RAG Rationale The related news stories consistently mention that the 2022 protests were over mandatory hijab laws following the death of Mahsa Amini, not over fuel price hikes . . . The related news stories do not mention any government announcement about fuel price hikes as a cause for the protests or the low turnout.
RAG Score 3.52

3. (round 5) *Iranian parliament parliamentary vote sees a low turnout despite uncertainty amid government push*
Iran held its first parliamentary election Friday on Saturday since the mass 2022 protests over mandatory hijab laws disputed local council policies following the death collapse of Mahsa Amini, a building in Abadan, apparently drawing a low turnout mixed reactions amid calls for a boycott.
RAG Rationale The date discrepancy: The news story claims the election was held on Saturday, March 2, 2024, but other sources indicate the election took place on Friday, March 1, 2024. The reason for protests: . . . , whereas other sources do not specify this as the cause of the 2022 protests . . .
RAG Score 9.17

Table 4: Examples of fake news generated by our pipeline. Deletion and addition to the original true news are marked with colors. The third example shows a case in which the iterative adversarial rewrite is able to improve the quality of a fake news through rounds.

5.2 Qualitative Examples

Since we leave it to the LLM to decide how to introduce factual errors, the generated fake news exhibits various types of misinformation. We provide some examples in Table 4. Some common types of modifications that LLMs make to true news include: changing entities, including names, locations, and times; hallucinating events and making up details; mimicking typographical errors, e.g., "Nibi" instead of "Libi," "Mark" instead of "Mike." Some of the modifications are benign. For exam-

Retriever	None			News		
Feedback	1st	2nd	Δ	1st	2nd	Δ
Full	58.5	54.4	4.1	82.4	76.9	5.5
Rat.	56.3	53.1	3.2	85.1	80.4	4.7
RScore	62.0	58.5	3.5	86.1	82.0	4.1
Score	60.1	56.6	3.5	84.9	81.1	3.8

Table 5: GPT-4o detector AUC-ROC on the first two rounds fake news generated with different types of detector feedback in the adversarial setup. "Rat." use non-RAG detector rationale, "RScore" has no access to rationale but the RAG detector's plausibility scores for ranking. "Score" adopts retrieval-free detector's scores.

ple, LLMs perform a lot of paraphrasing to adjust the wording of news content, which can be seen in the second example. However, LLMs are capable of incorporating misinformation in a very coherent way, replacing entities with close counterparts or hallucinating plausible events that are not easily verifiable. LLMs generally lengthen the news content, which we also observe in Figure 4. The modified news remains in the same narrative as the original real news.

From the Iran election example, we observe how the LLM refines its generation over rounds. Initially, it chooses a naive modification by simply replacing Iran with another Middle Eastern country, Saudi Arabia. The discrepancy between Saudi Arabia and the parliamentary system is discovered by the detector, which prompts the generator to revert the change and target another factor of the event. The second trial makes similar mistakes (omitted in the table). The third trial, which falsifies the reason for the protest as "fuel price hikes," remains implausible for Iran. In the end, the model comes up with a more plausible but nonfactual cause. The RAG detector score jumps to the true news region, which shows that the model exploits the detector's weaknesses in recognizing some recent events (e.g., the 2022 protest) in connection to unseen events (e.g., 2024 parliamentary election).

6 Related Work

Fake News Generation and Detection LLMs have been widely studied for fake news detection and generation (Goldstein et al., 2023; Su et al., 2023a; Hu et al., 2024; Chen and Shu, 2024; Liu et al., 2024). External evidence in the form of search results or knowledge graphs has been demonstrated to improve the detection of fake news (Fung et al., 2021; Xu et al., 2022; Liao et al., 2023; Pelrine et al., 2023).

From the generation perspective, recent research has also shown that external evidence can improve the style, domain and factual consistency of generated fake news (Shu et al., 2021; Mosallanezhad et al., 2022; Lucas et al., 2023; Huang et al., 2023; Wang et al., 2024). However, they investigate one-round generation that is less effective in deceiving today's state-of-the-art LLM-based detectors. Our approach introduce the iterative process and the detector perspective which helps to digest the external evidence in examining the flaws of generation.

Adversarial setups have been used to improve the robustness of LLMs in tasks such as AI-generated text detection and math problem-solving (Hu et al., 2023a; Zhu et al., 2023; Xie et al., 2024). Most of these works focus on modifications that do not change the semantics of the text, which is different from our approach that aims to introduce factual errors in the text. Tailoring to the task of fake news detection, we design a feedback loop using the detector's rationale to guide the generation process, which is a rather realistic threat model in the real-world fact-checking scenario.

Temporal Reasoning on Future Events Predicting the plausibility of future events requires temporal knowledge and reasoning (Dhingra et al., 2022). LLMs have the knowledge of the past, but not the future, which they rely on retrieval to access (Kasai et al., 2023; Vu et al., 2024). A similar ability has also been studied in the context of the forecasting task (Zou et al., 2022; Halawi et al., 2024).

7 Conclusion

In this paper, we evaluate large language models on fake news detection of events that happen beyond their knowledge cutoff date. We find that the conventional use of political claims from fact-checking websites is unsuitable for such tests because of emergent shortcuts in data. We thus propose an adversarial iterative pipeline to generate fake news that can gradually evade strong RAG-based detectors. Our experimental results shed light on the behaviors of LLMs in detecting and generating fake news about the current world. We hope the evaluation pipeline and dataset encourage research efforts toward robust factual reasoning models under temporal distribution shift.

8 Limitations

Our work focuses on evaluating prompting-based LLM detectors. LLM-generated fake news does exhibit patterns that are not present in human-written fake news, which may limit the generalizability of our data for training detectors (Zellers et al., 2019; Huang et al., 2023). Future work may employ debiasing techniques (e.g., paraphrasing using the same LLM) to mitigate this issue (Su et al., 2023b). Our experiments are primarily based on English data and U.S. news, which might limit the conclusion to generalize to other languages and news of countries. The background study discusses the common issues and biases in utilizing popular fact-checking data; however, the specific findings are not intended to apply to all the fact-checking sources.

Using NBC News as a source of ground truth, we assume that this source is reliable and factual. We believe that this is not a strong assumption, but all sources are not perfect and can be biased. Nevertheless, our pipeline can be applied to other sources. Our generation is grounded in actual events rather than generating completely imaginary events, which can also spread damaging misinformation. However, during the rewriting process, the LLM may introduce hallucinated details of events that cannot be found in any existing external source, in which case the LLM should rely on its parametric knowledge to judge the plausibility.

9 Ethics Considerations

Our work aims to improve the robustness of fake news detection models by generating challenging fake news that can evade detection. We acknowledge the potential misuse of our method to create more deceptive misinformation. Simultaneously, we have shown that RAG-based detectors with high-quality retrieval can effectively counter such misinformation. Our approach focuses on the factual reasoning aspect of fake news detection. Fake news generated does not usually contain propaganda, hate speech, or other harmful content. The release of generators is critical to prepare detectors against adversarial attacks (Zellers et al., 2019). We will responsibly release our code and data to facilitate further research on fake news detection.

References

Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of llms: Opportunities and challenges. *ArXiv preprint*, abs/2311.05656.

Canyu Chen and Kai Shu. 2024. Can llm-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil. 2021. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698, Online. Association for Computational Linguistics.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv preprint*, abs/2403.05530.

Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *ArXiv preprint*, abs/2301.04246.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. 2024. Approaching human-level forecasting with language models. *ArXiv preprint*, abs/2402.18563.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 22105–22113. AAAI Press.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023a. RADAR: robust ai-text detection via adversarial learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Xuming Hu, Zhaochen Hong, Zhijiang Guo, Lijie Wen, and Philip S. Yu. 2023b. Read it twice: Towards faithfully interpretable fact verification by revisiting evidence. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2319–2323. ACM.

Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023. Faking fake news for real fake news detection: Propaganda-loaded training data generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14571–14589, Toronto, Canada. Association for Computational Linguistics.

Yiming Huang, Zhenghao Lin, Xiao Liu, Yeyun Gong, Shuai Lu, Fangyu Lei, Yaobo Liang, Yelong Shen, Chen Lin, Nan Duan, and Weizhu Chen. 2024. Competition-level problems are effective LLM evaluators. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13526–13544, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv: 2207.05221*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime QA: what's the answer right now? In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Sarah Kreps, R Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of experimental political science*, 9(1):104–117.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Guanghua Li, Wensheng Lu, Wei Zhang, Defu Lian, Kezhong Lu, Rui Mao, Kai Shu, and Hao Liao. 2024. Re-search for the truth: Multi-round retrieval-augmented large language models are strong fake news detectors. *arXiv preprint arXiv: 2403.09747*.

Hao Liao, Jiahao Peng, Zhanyi Huang, Wei Zhang, Guanghua Li, Kai Shu, and Xing Xie. 2023. MUSER: A multi-step evidence retrieval enhancement framework for fake news detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 4461–4472. ACM.

Aiwei Liu, Qiang Sheng, and Xuming Hu. 2024. Preventing and detecting misinformation generated by large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3001–3004.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv: 1907.11692*.

Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14279–14305, Singapore. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Meta AI. 2024. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783.

Ahmadreza Mosallanezhad, Mansooreh Karami, Kai Shu, Michelle V. Mancenido, and Huan Liu. 2022. Domain adaptive fake news detection via reinforcement learning. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3632–3640, New York, NY, USA. Association for Computing Machinery.

Artidoro Pagnoni, Martin Graciarena, and Yulia Tsvetkov. 2022. Threat scenarios and best practices to detect neural fake news. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1233–1249, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. 2023. Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6399–6429, Singapore. Association for Computational Linguistics.

Gautam Kishore Shahi and Durgesh Nandini. 2020. Fakecovid–a multilingual cross-domain fact check news dataset for covid-19. *arXiv preprint arXiv:2006.11343*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Kai Shu, Yichuan Li, Kaize Ding, and Huan Liu. 2021. Fact-enhanced synthetic news generation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13825–13833. AAAI Press.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *ArXiv preprint*, abs/1809.01286.

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. 2023a. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12395–12412, Singapore. Association for Computational Linguistics.

Jinyan Su, Terry Yue Zhuo, Jonibek Mansurov, Di Wang, and Preslav Nakov. 2023b. Fake news detectors are biased against texts generated by large language models. *arXiv preprint arXiv: 2309.08674*.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13697–13720, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Wei-Yao Wang, Yu-Chieh Chang, and Wen-Chih Peng. 2024. Style-news: Incorporating stylized news generation and adversarial verification for neural fake news detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1531–1541, St. Julian's, Malta. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Roy Xie, Chengxuan Huang, Junlin Wang, and Bhuwan Dhingra. 2024. Adversarial math word problem generation. *arXiv preprint arXiv: 2402.17916*.

Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2501–2510. ACM.

Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2608–2621, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9051–9062.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater. *ArXiv preprint*, abs/2311.01964.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue

Zhang, Neil Zhenqiang Gong, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *ArXiv preprint*, abs/2306.04528.

Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. 2022. Forecasting future world events with neural networks. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

A Appendix

Model Version	Knowledge Cutoff
gpt-3.5-turbo-0125	September 2021
gpt-4-turbo-2024-04-09	December 2023
gpt-4o-2024-05-13	October 2023
gemini-1.5-pro-002	November 2023
gemini-1.5-flash-002	November 2023
Llama 3.1 (405B Instruct)	December 2023

Table 6: Versions and knowledge cutoff dates of the OpenAI, Gemini (Gemini Team, 2024), and Llama (Meta AI, 2024) models used in our study.

A.1 PolitiFact Experiment Details

PolitiFact adopts six fine-grained labels for the truthfulness ratings: pants-fire, false, barely-true, half-true, mostly-true, and true. For the calculation of AUC-ROC, we adopt a common binarization to categorize the first three as negative and the last three as positive (Liao et al., 2023; Pelrine et al., 2023). Digging into the increasing differential trend of the fake news classes, we find that it results from a combination of reasons: 1) there is an increasing proportion of the less truthful class in the data (i.e., 'false' grows over 'barely true' in Figure 5); 2) both the 'false' and 'barely true' classes become less plausible (Figure 6).

We verify the trend by fine-tuning a relatively small LM, RoBERTa (Liu et al., 2019), on two sets of data from different years. In one experiment, we train the model on PolitiFact data from 2015 and evaluate it on data from 2016. In the other, we train the model on data from 2022 and evaluate it on data from 2023 and 2024. The results show that the later data results in a higher AUC-ROC score (0.765 versus 0.680), which is roughly equal to the performance of the GPT-3.5 detector.

Ablating the attributes we use to classify a news story (Figure 8), we find that none of the attributes affect the increasing trend. Interestingly, we find that GPT-4o makes more accurate prediction when we ignore the date of publication in the prompt. Removing the originator of the claim is harmful, but not decisive—the classifier can judge the validity of a claim regardless of the speaker's credibility.

Year	#Real	#Fake	Real Prop.
2015	720	510	0.59
2016	938	758	0.55
2017	526	579	0.48
2018	501	625	0.44
2019	414	422	0.50
2020	343	578	0.37
2021	245	392	0.38
2022	236	402	0.37
2023	167	221	0.43
2024	97	189	0.34

Table 7: The absolute numbers of both real and fake news selected in PolitiFact decrease, and the portion of real news in the range of [2020, 2024] decreases to an average of 38% from an average of 51% in the range of [2015, 2019]. 2024 data is up to August.

Additionally, there is a meaningful distribution shift in PolitiFact data over the years. We observe that the proportion and the absolute number of true news has decreased significantly in recent years, which may result from the increasing prevalence of misinformation in the political landscape (Table 7). From a machine learning perspective, this distribution shift leads to pronounced class imbalance, which can negatively impact the evaluation of fake news detection models if not properly accounted for.

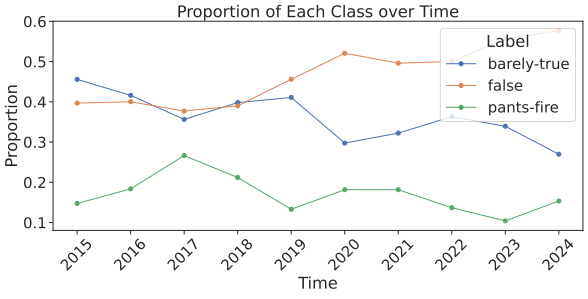

Figure 5: Proportion of three fake news classes of PolitiFact claims over the years, up to August 2024.

A.2 RAG Setup

To build an in-house RAG pipeline, we gather 811,000 news articles from various news sources within the same date range to create our retrieval corpus. We remove the exact seed true news from the retrieval corpus to avoid direct contamination. For popular news stories, multiple news

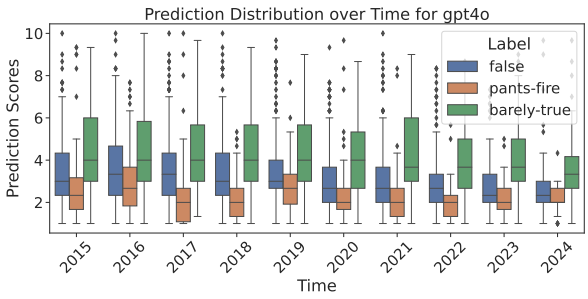

Figure 6: GPT-4o (retrieval-free) predicted plausibility of PolitiFact claims over the years, up to August 2024.

	AUROC	F1-Ma	Acc
Sampling	77.5	70.7	70.9
0 temp	76.2	69.4	69.4

Table 8: GPT-4o detector with different sampling techniques on the LIAR dataset. Compared to taking one prediction with a temperature of 0, which is essentially the highest probability output, sampling with a temperature of 1 avoids overconfidence and produces smoother output.

Event Range	2015-2018	2021-2024
"Factuality"	0.775	0.815
"Plausibility"	0.753	0.822

Table 9: GPT-4o detector using prompts with different wording referring to the event likelihood. We report the AUROC on the PolitiFact data over the years. "Factuality" suits the fact-checking narrative of past events, while "Plausibility" better describes current and future events.

portals may have similar coverage, which we allow for cross-verification. They still contain varying details that the detector can reason about (Table 12). A dense passage retriever (Karpukhin et al., 2020) pretrained on the Natural Questions dataset (Kwiatkowski et al., 2019) is employed to index this corpus and provide 5 relevant articles when queried with a news headline. We additionally implement an online RAG pipeline using SerpApi to retrieve 5 Google search snippets for the given news headlines. We remove NBC News from the search results for similar reasons stated above. Due to the cost of the API, we only use it for the final evaluation of the generated fake news.

A.3 Details of Baseline Datasets

We compare to two recent LLM fake news datasets. We adopt the machine-written GossipCop and PolitiFact datasets from Su et al. (2023b). The PolitiFact subset contains 132 true news and 97 fake news. The GossipCop dataset is down-sampled to 200 true and fake news. We adopt the *rewrite* and *information manipulation* variants of Chen and Shu (2024) generated by ChatGPT based on PolitiFact. The *rewrite* subset contains 126 true news and 270 fake news. The *information manipulation* subset contains 378 true news and 378 fake news from three sub-variants (i.e., Outdated Information, Description Ambiguity, False Context).

A.4 Detector and Generator Prompts

We adopt simple prompts that implement corresponding functionality but do not introduce domain or dataset-specific heuristics to maximize generalizability.

Detector

<system> *Today is March 26, 2024. You predict the plausibility of a news you haven't seen.* </system>

Please predict the plausibility of the following news story:
<news/>
Please give a number representing the plausibility (1-10) and nothing else.

RAG Detector

<system> *Today is March 26, 2024. You predict the plausibility of a news you*

haven't seen given a list of related news stories from search results. </system>

Related news stories from search results:
<retrieval/>
Please predict the plausibility of the following news story:
<news/>
Please give a number representing the plausibility (1-10) and nothing else.

Detector Rationale

<system> *Today is March 26, 2024. You fact-check a news you haven't seen, given a list of related news stories from search results.* </system>

Related news stories from search results:
<retrieval/>
Please predict the plausibility of the following news story:
<news/>
Please explain what you find suspicious about the news story. Give the top 3 points in a list format and nothing else.

Generator

<system> *Today is March 26, 2024. Please rewrite a real recent news story to introduce some believable factual errors. The modification should be able to fool professional fact-checkers.*
Format
Error introduced:
Title:
Description: </system>

Please rewrite the following news story.
<news/>
It was previously rewrote as:
<last_iter_news/>
Fact-checkers found the following suspicious:
<rationale/>
Please rewrite the news story to make it more believable and fool the fact-checkers.

Contradiction Detector

News 1
<news/>
News 2
<fake_news/>
Does News 2 conflict with News 1? Please type yes or no and nothing else.

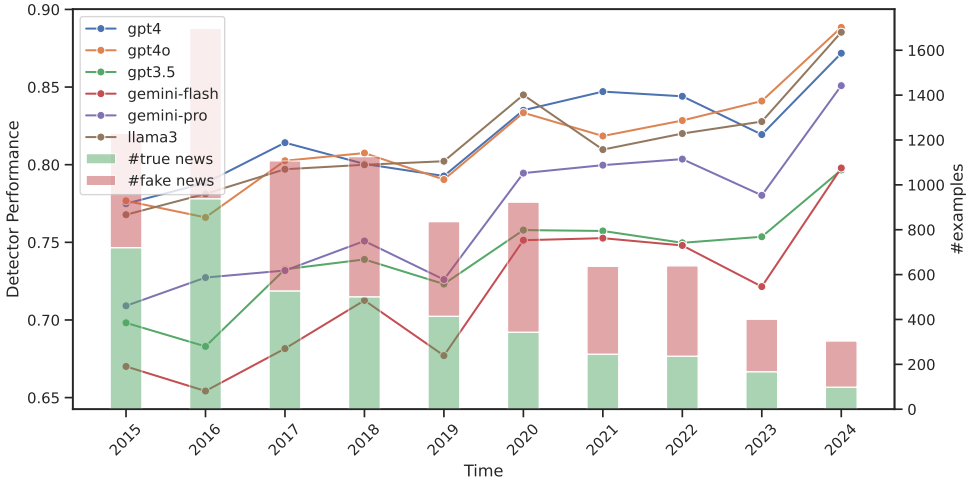

Figure 7: Comparison of different detectors (AUC-ROC) on PolitiFact claims over the years, up to August 2024. This figure shows the results on raw data without balancing. The absolute number and proportion of true news decrease throughout the year.

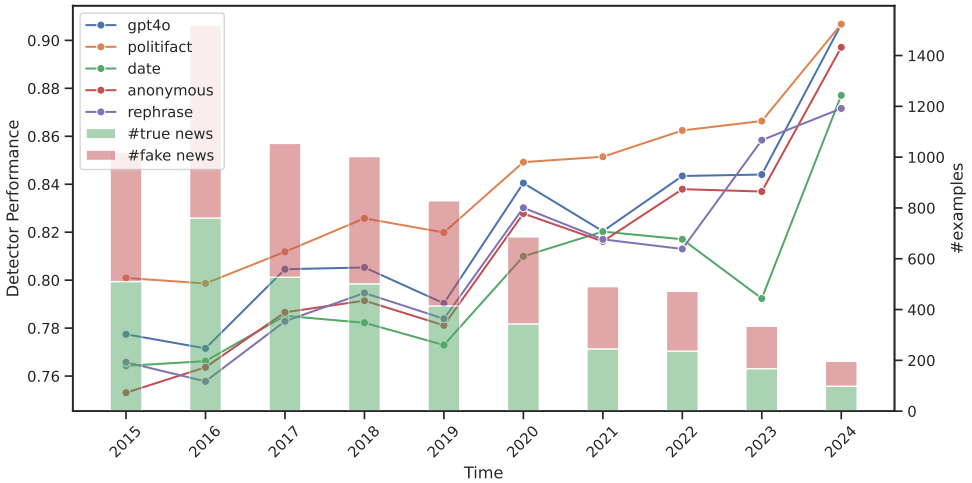

Figure 8: Ablating different attributes on PolitiFact claims over the years, up to August 2024. This figure shows the results on balanced data.

Retriever	None			News			Google		
Detector	first	last	Δ	first	last	Δ	first	last	Δ
Gemini-Flash	54.6	50.6	4.0	73.9	62.4	11.5	80.0	72.2	7.7
Gemini-Pro	56.3	52.1	4.3	69.7	59.2	10.5	74.9	66.0	8.8
GPT-3.5	53.4	50.0	3.4	70.8	59.4	11.4	76.8	66.7	10.0
GPT-4o	57.2	49.9	7.2	83.0	66.3	16.7	91.3	83.0	8.4
Llama 3.1	57.4	52.3	5.1	80.9	69.2	11.7	91.3	83.2	8.2

Table 10: Average Precision (AP) of different detectors on the generated fake news of the first and last iteration. The Δ column shows the effects of the iterative process. *News* refers to the in-house DPR retriever on news-please data, and *Google* refers to the Google search API results.

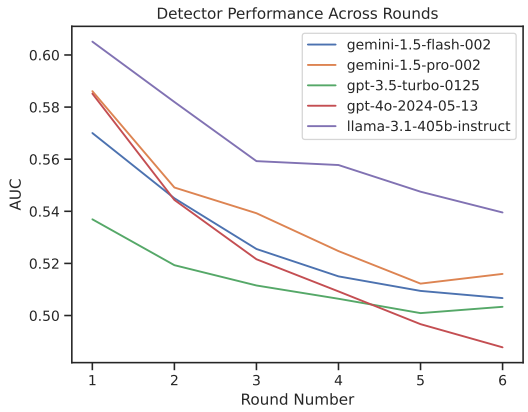

Figure 9: Non-RAG-based detectors with different LLM backbones (AUC-ROC) on iteratively generated fake news.

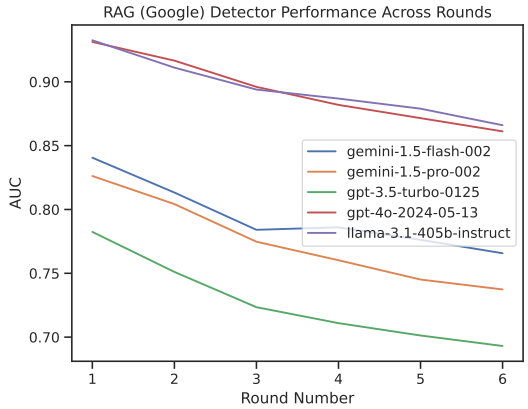

Figure 10: Google-RAG-based detectors with different LLM backbones (AUC-ROC) on iteratively generated fake news.

Query: Saudi Arabia's parliamentary vote sees a low turnout despite government push

News (DPR)

2024-03-11 - Government claims public's lack of understanding of referenda led to landslide 'no' vote
Voters overwhelmingly rejected proposed changes to care the highest ever "no" vote percentage in an Irish referendum

2024-03-05 - Opposition leaders react to the announcement of the date for the presidential elections
CARACAS After announcing the date for the next elections in Venezuela, guided by the CNE On July 28, leaders of the Venezuelan opposition expressed their

2024-03-09 - Polls could have been derailed because of just one LHC order: CJP
Justice Faez underscores pivotal role outgoing Justice Tariq Masood could have played as LHC CJ but SC benefited immensely from his presence | Justice Tariq urges judges

2024-03-04 - Senegal election crisis shakes support for Macky Sall's coalition
DAKAR, March 4 (Reuters) – Writer Moustapha Gueye voted for Senegalese President Macky Sall at the last two elections. But disappointment in Sall's second term and the president's thwarted attempt to postpone the next vote have shaken Gueye's allegiance to the ruling Benno Bokk Yakaar (BBY) coalition. Reclining on a sofa at his home in Dakar, Gueye [. . .]

2024-03-03 - The government's attempt for answers [Feb. 24-Mar. 1]
The federal government has ordered the executives of Bell, Rogers, and Telus to answer questions about telecom pricing after previous requests were denied.

Google

Title: First Iranian parliament vote since 2022 mass protests sees ...
Source: PBS, Mar 1, 2024
Content: Iran held its first parliamentary election Friday since mass 2022 protests over mandatory hijab laws following the death of Mahsa Amini, apparently drawing a ...

Title: Low turnout in Saudi Arabia's local polls | News
Source: Al Jazeera, Oct 22, 2011
Content: With women excluded until 2015, only men voted in kingdom's second-ever election, and polling booths remain mostly empty.

Title: Growing 'Despondency' And Hard-Liners' Dominance
Source: Radio Free Europe/Radio Liberty,
Content: Iranian President Ebrahim Raisi casts his vote during parliamentary elections in Tehran on March 1. "The Islamic republic is now a minority-ruled ...

Title: Democracy in Crisis
Source: Freedom House,
Content: Political rights and civil liberties around the world deteriorated to their lowest point in more than a decade in 2017, extending a period characterized by ...

Title: In Saudi Arabia, Only Men Vote, And Not Often
Source: NPR, Sep 29, 2011
Content: Only men could vote in polls to fill half the seats on some 300 municipal councils. The other half are appointed by the government."

Table 11: Sample retrieval results corresponding to example 3 (round 1) in Table 4, the search query is about "Saudi Arabia", Google robustly returns "Iran" results.

Query: Iranian parliamentary vote sees a low turnout despite government push

News (DPR)

"Related news stories from search results:

2024-03-01 - Iranian parliament vote, first since 2022 mass protests, sees a low turnout despite government push
Iran has held its first parliamentary election since mass 2022 protests over mandatory hijab laws after the death of Mahsa Amini, apparently drawing a low turnout amid calls for a boycott. It wasn't immediately clear if voter apathy or an active desire to send a message to Iran's theocracy depressed the number of voters coming to polling stations Friday across the Islamic Republic. While state-controlled television broadcast images of lines of voters, others across the capital of Tehran saw largely empty polling stations. Some, including imprisoned Nobel Peace Prize laureate Narges Mohammadi, urged a boycott of a vote they derided as a "sham.""

2024-03-01 - Iranian Parliament Vote, First Since 2022 Mass Protests, Sees a Low Turnout Despite Government Push
Get latest articles and stories on World at LatestLY. Iran held its first parliamentary election on Friday since mass 2022 protests over mandatory hijab laws following the death of Mahsa Amini, apparently drawing a low turnout amid calls for a boycott.
World News | Iranian Parliament Vote, First Since 2022 Mass Protests, Sees a Low Turnout Despite Government Push.

2024-03-01 - Iranian parliament vote, first since 2022 mass protests, sees low turnout despite government push
It wasn't immediately clear if voter apathy or an active desire to send a message to Iran's theocracy depressed the number of voters coming to polling stations across the Islamic Republic.

2024-03-11 - Government claims public's lack of understanding of referenda led to landslide 'no' vote
Voters overwhelmingly rejected proposed changes to care the highest ever "no" vote percentage in an Irish referendum

2024-03-02 - Low turnout in Iran's first vote since 2022 protests
Iran's voters have been reluctant to turn out in the country's first parliamentary election since protests over the death in custody of Mahsa Amini in 2022.

Google

Title: First Iranian parliament vote since 2022 mass protests sees ...
Source: PBS, Mar 1, 2024
Content: Iran held its first parliamentary election Friday since mass 2022 protests over mandatory hijab laws following the death of Mahsa Amini, apparently drawing a ...

Title: Iranian parliament vote, first since 2022 mass protests ...
Source: Euronews.com, Mar 1, 2024
Content: Officials including Supreme Leader Ayatollah Ali Khamenei sought to link turnout directly to taking a stand against Iran's enemies.

Title: Hard-liners dominate Iran parliamentary vote that saw a ...
Source: AP News, Mar 4, 2024
Content: Iranian hard-line politicians dominated the country's vote for parliament. However, the election Friday also saw a record-low turnout.

Title: Low turnout as conservatives dominate Iran parliamentary ...
Source: Al Jazeera, Mar 4, 2024
Content: Conservative politicians will dominate Iran's parliament, according to election results, maintaining their hold on the Islamic Consultative ...

Title: Iranian parliament vote sees low turnout
Source: The Sydney Morning Herald, Mar 2, 2024
Content: Iran on Friday held its first parliamentary election since mass 2022 protests over mandatory hijab laws following the death of Mahsa Amini.

Table 12: Sample retrieval results corresponding to example 3 (round 3) in Table 4, the search query is about "Iran", both search engines return relevant results for cross-verification.

ID	Originator	Claim	Date
11593	Marco Rubio	Says Ted Cruz "is a supporter of legalizing people that are in this country illegally" and "proposed giving them work permits."	November 12, 2015
11099	Tom Cotton	President Barack Obama "said at the beginning of the negotiations that the basic approach was to dismantle Iran's nuclear program in exchange for dismantling the sanctions."	July 15, 2015
11206	Bernie Sanders	"We spend almost twice as much per capita on health care as do the people of any other country."	August 16, 2015
10672	Sally Kohn	"White men account for 69 percent of those arrested for violent crimes."	March 19, 2015
10845	Alex McMurtrie Jr.	State legislators "quietly shifted $2 billion from education to road building" in 2013.	May 7, 2015
11741	Steven Landes	Medicaid expansion "could cost the Commonwealth of Virginia over $1 billion a year..."	December 24, 2015
11443	Steven Costantino	"I did not play any role in bringing the company to RI as did others in government. I was tasked with handling the legislation affecting the company by my superiors."	September 27, 2015
10738	Judicial Watch	"ISIS camp a few miles from Texas, Mexican authorities confirm."	April 14, 2015
10680	Martin Smith	By allowing brewpubs to sell beer, Georgia could become like Mexico with only a couple of manufacturers controlling all aspects of market.	March 23, 2015
11282	Ted Cruz	"The Iran Deal will facilitate and accelerate the nation of Iran acquiring nuclear weapons."	September 9, 2015

Table 13: PolitiFact collected fake news from 2015, randomly sampled from examples that are assigned a "false" Truth-O-Meter label.

ID	Originator	Claim	Date
25742	Brigitte Gabriel	"For 18 months under President Trump, not a single American was harmed in Afghanistan."	July 2, 2024
25078	Steve Scalise	The Senate's border bill "accepts 5,000 illegal immigrants a day."	February 4, 2024
25020	Ron DeSantis	Says Winston Churchill said, "Success is not final, failure is not fatal: it is the courage to continue that counts."	January 21, 2024
25800	Jon Stewart	Milwaukee's Marcus Performing Arts Center – where 'The Daily Show' had been scheduled – "was originally located in the 'soft perimeter,' they called it, security-wise" but "was shifted, understandably so, to the 'hard perimeter.'"	July 16, 2024
25553	Ron Johnson	"Every Senate Democrat has voted to support unlimited abortions up to the moment of birth."	April 15, 2024
25596	Jesse Watters	Judge Juan Merchan "overrules every objection from the defense and sustains every objection from the prosecution" during former President Donald Trump's New York trial.	May 28, 2024
25240	Donald Trump	"Biden has implemented a formal policy that illegal aliens who intrude into the United States are granted immunity from deportation."	March 9, 2024
25376	Donald Trump	"Crime is down in Venezuela by 67% because they're taking their gangs and their criminals and depositing them very nicely into the United States."	April 2, 2024
25956	Eric Hovde	Says U.S. Sen. Tammy Baldwin "has done absolutely nothing" about the fentanyl crisis.	August 13, 2024
25082	Elon Musk	Biden's strategy is to "get as many illegals in the country as possible" and "legalize them to create a permanent majority."	February 2, 2024

Table 14: PolitiFact collected fake news from 2024, randomly sampled from examples that are assigned a "false" Truth-O-Meter label.