TripleFact: Defending Data Contamination in the Evaluation of LLM-driven Fake News Detection

Cheng Xu[1,3] **Nan Yan**[2,3]
[1] University College Dublin [2] Georgia Institute of Technology [3] Bebxy
cheng.xu1@ucdconnect.ie nan.yan@gatech.edu

Abstract

The proliferation of large language models (LLMs) has introduced unprecedented challenges in fake news detection due to benchmark data contamination (BDC), where evaluation benchmarks are inadvertently memorized during the pre-training, leading to the inflated performance metrics. Traditional evaluation paradigms, reliant on static datasets and closed-world assumptions, fail to account the BDC risk in large-scale pre-training of current LLMs. This paper introduces **TripleFact**[1], a novel evaluation framework for fake news detection task, which designed to mitigate BDC risk while prioritizing real-world applicability. TripleFact integrates three components: (1) Human-Adversarial Preference Testing (**HAPT**) to assess robustness against human-crafted misinformation, (2) Real-Time Web Agent with Asynchronous Validation (**RTW-AV**) to evaluate temporal generalization using dynamically sourced claims, and (3) Entity-Controlled Virtual Environment (**ECVE**) to eliminate entity-specific biases. Through experiments on 17 state-of-the-art LLMs, including GPT, LLaMA, and DeepSeek variants, TripleFact demonstrates superior contamination resistance compared to traditional benchmarks. Results reveal that BDC artificially inflates performance by up to 23% in conventional evaluations, while TripleFact Score (**TFS**) remain stable within 4% absolute error under controlled contamination. The framework's ability to disentangle genuine detection capabilities from memorization artifacts underscores its potential as a fake news detection benchmark for the LLM era.

1 Introduction

The proliferation of misinformation and disinformation in the digital age has made fake news detection a critical challenge for society, threatening democratic processes, public health, and social stability (Allcott and Gentzkow, 2017; Shu et al., 2017; Zafarani et al., 2019; Rocha et al., 2021; Lewandowsky, 2025). Traditional approaches to this task have relied on supervised machine learning frameworks, where models are trained on labeled datasets to classify news articles as "fake" or "real" based on like linguistic patterns, source credibility, or network propagation features (Pérez-Rosas et al., 2018). These methods, often built on classical machine learning models, e.g. SVM (Cortes and Vapnik, 1995), LSTM (Hochreiter and Schmidhuber, 1997), Random Forest (Xu et al., 2022), operated under the assumption of a clear separation between training and testing data, a paradigm now disrupted by the advent of large language models (LLMs), such as GPT-4 (OpenAI, 2024), LLaMA (Touvron et al., 2023a) and DeepSeek (DeepSeek-AI et al., 2024).

The rise of LLMs, pre-trained on vast corpora spanning diverse domains and timelines, has introduced unprecedented challenges for benchmark-driven evaluation (Wu et al., 2024; Papageorgiou et al., 2024). Unlike traditional models, LLMs are exposed to trillions of tokens during pre-training (OpenAI, 2024; GLM et al., 2024), often including datasets later used for evaluation. This benchmark data contamination (BDC), where test examples or related information are inadvertently included in pre-training data—renders conventional train-test splits ineffective, as models may already "memorize" benchmark-specific patterns. While this issue affects many NLP tasks, fake news detection is uniquely vulnerable due to its reliance on real-world, time-sensitive claims and the propensity of LLMs to internalize factual and counterfactual information alike (Xu et al., 2024). For instance, an LLM trained on historical news archives may recognize a debunked conspiracy theory as a known pattern, artificially inflating its detection performance without genuine understanding.

Recent studies have highlighted the BDC risks in LLM evaluations (Törnberg, 2023), yet little work

[1] https://github.com/chengxuphd/triplefact

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8808–8823

has addressed its implications for fake news detection task. Current practices, such as zero-shot testing without fine-tuning, fail to account for the temporal and contextual dynamics of misinformation (Pelrine et al., 2023). For example, LLMs pre-trained on data up to 2023 cannot reliably detect fake news emerging in 2024, but their performance on older benchmarks may still be overstated due to contamination (Horne et al., 2019). This creates a critical gap between evaluation benchmarks and real-world applicability.

In this paper, we argue that the fake news detection community must urgently re-evaluate its methodologies to align with the realities of the LLM era. We critique the limitations of current evaluation practices, demonstrate how data contamination skews performance metrics, and propose a framework prioritizing temporal robustness, adversarial generalization, and contamination-aware evaluation. Our **contributions** are threefold:

1. We systematize the differences between traditional and LLM-driven fake news detection, emphasizing the inadequacy of conventional benchmarks.

2. We provide empirical evidence of BDC's impact on fake news detection task using case studies from popular LLMs.

3. We introduce TripleFact, a novel evaluation framework that mitigates contamination risks while maintaining practical relevance for real-world deployment.

The remainder of this paper is structured as follows: Section 2 contrasts traditional and LLM-based approaches to fake news detection, analyzes the data contamination problem and its consequences. Section 3 presents our proposed framework, followed by the experiments and analysis in Section 4.

2 The Benchmark Data Contamination Problem in Fake News Detection

This section contrasts the differences between traditional and LLMs-driven fake news detection, and explains why BDC exists only in LLM-driven fake news detection. Subsequently, we formalizes the concept of BDC within the framework of LLM evaluation, distinguishes its scope from broader domain adaptation challenges, and clarifies its exclusive relevance to benchmarking contexts. Finally, we review the current state of research on BDC.

Aspect	Traditional Approach	LLM-Driven Approach
Data Scope	Curated, domain-specific datasets	Massive, open-domain corpora
Training	Supervised fine-tuning	pre-training on diverse text
Evaluation	Closed-world train-test splits	Zero/few-shot prompting
Key Limitation	Limited generalization	Data contamination risks

Table 1: Comparison of Traditional vs. LLM-Driven Fake News Detection

2.1 Traditional Paradigm vs. LLM Paradigm

Traditional fake news detection systems relied on supervised learning, where models were trained on labeled datasets to classify news item based on linguistic, stylistic, or social network features (Zhou and Zafarani, 2020). Key characteristics include feature engineering (Pérez-Rosas et al., 2018), using classical machine learning models (Xu and Kechadi, 2023), and closed-world evaluation (Shu et al., 2020). For example, the LIAR dataset (Wang, 2017) became a gold standard for evaluating models on political fact-checking, with its clean train-test splits and labeled credibility ratings. However, such benchmarks assumed closed-world evaluation, where test data was both unseen and representative of real-world distribution—a premise challenged by LLMs' pre-training on open-world corpora.

The advent of LLMs like BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020) shifted fake news detection from feature engineering to leveraging knowledge embedded in pre-trained models. LLMs ingest trillions of tokens from diverse sources, e.g., common crawl, Wikipedia[2], social media (Xu and Yan, 2023), potentially including benchmark-related information used in traditional evaluation (Liu et al., 2024). As for the evaluation, more common practice is not fine-tuning, researchers directly prompt LLMs to classify the news based on their truthfulness, relying on their pre-trained knowledge (Brown et al., 2020; Pelrine et al., 2023). Test examples may appear in pre-training data, leading to inflated performance metrics. For instance, if GPT-4 was pre-trained on LIAR's test set, its zero-shot accuracy becomes unreliable (Razeghi et al., 2022).

Fake news detection task is uniquely affected by BDC issues due to its need for massive contextual information, which we summarized in the following main three factors:

- **Memorization of Claims:** LLMs internalize both factual and false claims during pre-training. For example, GPT-3 can regurgitate

[2] https://wikipedia.org/

debunked conspiracy theories verbatim (Carlini et al., 2021), conflating memorization with detection capability.

- **Temporal Misalignment:** Fake news evolves rapidly, but LLMs are trained on historical data. A model pre-trained on 2021 data cannot reliably detect 2023 misinformation, yet contamination from older benchmarks (e.g., LIAR) may still inflate perceived performance (Stepanova and Ross, 2023).

- **Domain Shifts:** LLMs pre-trained on general corpora lack fine-grained signals (e.g., emerging slang, niche disinformation tactics), leading to overconfidence in misclassified examples (Wen et al., 2024).

As summarized in Table 1, the transition from traditional machine learning models to LLM-driven approaches has fundamentally altered the landscape of fake news detection. While traditional models relied on carefully curated features and benchmarks, they are limited by assumptions of a closed-world evaluation setting, which are increasingly invalidated by the open-world nature of LLMs. LLMs, though powerful, pose new challenges such as BDC and memorization, which can artificially inflate their performance on well-known datasets. Fake news detection task, in particular, suffers from these issues due to its dependence on up-to-date and vast contextual information, making it an area ripe for further research and innovation to address these limitations effectively.

2.2 Formal Definition of Benchmark Data Contamination (BDC)

Benchmark Data Contamination refers to the phenomenon where evaluation benchmark data, including test splits, metadata, related context, or task-specific patterns, are inadvertently included in the pre-training corpus of LLMs, leading to inflated or unreliable performance metrics during evaluation. As articulated in Xu et al. (2024), BDC arises when LLMs encounter benchmark-related information during pre-training, enabling them to "memorize" answers or patterns rather than demonstrating genuine task-solving capabilities. This undermines the validity of evaluation protocols, as models may exploit memorized data instead of generalizing to unseen examples.

Critically, BDC is not synonymous with domain adaptation or general overfitting. Its relevance is strictly tied to the evaluation paradigm. BDC manifests only when models are evaluated on datasets that overlap with their pre-training data. For example, GPT-4's high performance on MMLU (Hendrycks et al., 2021) is compromised if test questions were present in its training corpus. As to the domain specific applications, when LLMs are deployed for real-world tasks (e.g., medical diagnosis, legal analysis), the absence of predefined benchmarks negates BDC concerns. Domain-specific performance depends on generalization, not contamination. This distinction is vital to avoid conflating BDC with legitimate domain adaptation. For instance, an LLM pre-trained on legal texts and fine-tuned for contract analysis does not suffer from BDC, as its training data are intentionally curated for the target domain. Thus, BDC is a technical challenge specific to evaluation integrity, not a blanket critique of LLM training practices.

2.3 Detection and Mitigation of BDC

The current research conducted on the BDC problem is broadly divided into two categories, detection and mitigation (Xu et al., 2024). The detection of BDC relies on methodologies that identify overlaps between pre-training corpora and evaluation benchmarks. Matching-based approaches, such as n-gram overlap analysis and embedding similarity checks, are foundational for identifying explicit contamination, where test examples appear verbatim in training data (Ippolito et al., 2023; Jiang et al., 2024; Li and Flanigan, 2024; Shi et al., 2024). For instance, Razeghi et al. (2022) demonstrated that models pre-trained on datasets containing benchmark questions exhibit inflated performance due to memorization. However, implicit contamination—where models internalize task-specific patterns rather than exact text—requires more sophisticated methods. Techniques like CDD (Contamination Detection via output Distribution) analyze the "peakedness" of LLM confidence scores to detect memorization, achieving 21.8–30.2% improvements over traditional methods in identifying subtle contamination (Dong et al., 2024; Deng et al., 2024). TS-Guessing (Testset Slot Guessing) further quantifies contamination by masking parts of benchmark questions and measuring reconstruction accuracy, revealing that GPT-4 achieves 57% exact matches in contaminated MMLU examples (Deng et al., 2023).

Mitigation strategies address BDC by redesigning evaluation protocols and curating uncontami-

nated datasets (Sun et al., 2025). Adversarial benchmarks like LatestEval dynamically generate test questions from recent texts, bypassing historical data dependencies (Li et al., 2024c). Regenerating benchmarks through paraphrasing or synthetic data creation also minimizes overlap while preserving task integrity (Xia et al., 2024; Zhu et al., 2024b,a; Ying et al., 2024). For models already exposed to contaminated data, post-hoc correction frameworks like TED (Trustworthy Evaluation via output Distribution) recalibrate confidence scores to mitigate inflated performance, reducing contamination-induced accuracy gains by up to 66.9% (Dong et al., 2024). Benchmark-free evaluation paradigms, such as human-in-the-loop fact-checking or real-time deployment metrics, further sidestep contamination by prioritizing real-world generalization over static benchmarks (Li et al., 2024b; Chiang et al., 2024; Yu et al., 2024).

In summary, BDC detection requires hybrid methods combining text matching, distribution analysis, and temporal validation, while mitigation hinges on dynamic evaluation design and transparency in training data curation. As emphasized by Xu et al. (2024) and Deng et al. (2024), resolving BDC demands collaborative efforts to standardize contamination audits, adopt time-sensitive benchmarks, and develop domain-specific evaluation frameworks that reflect real-world LLM applications rather than static benchmarks. These strategies collectively ensure that LLM evaluations measure genuine reasoning capabilities, not memorization artifacts.

3 Toward a Realistic Framework for LLM-Based Fake News Detection

The proliferation of data contamination in LLM evaluations necessitates a paradigm shift in assessing fake news detection systems. Existing frameworks often rely on static benchmarks or computationally intensive synthetic data generation (Yu et al., 2024; Zhu et al., 2024b; Xia et al., 2024), failing to address real-world dynamics while incurring prohibitive resource costs. To bridge this gap, we propose **TripleFact**, a novel evaluation framework that combines human-adversarial testing, real-time web validation, and entity-debiased virtual environments as shown in Figure 1. The TripleFact framework addresses the dual challenges of data contamination and computational inefficiency in LLM-based fake news detection by integrating three synergistic components: Human-Adversarial Preference Testing (**HAPT**), Real-Time Web Agent with Asynchronous Validation (**RTW-AV**), and Entity-Controlled Virtual Environment (**ECVE**). This triad approach prioritizes contamination resistance, low computational overhead, and practical relevance, addressing limitations inherent to traditional methodologies.

3.1 Component 1: Human-Adversarial Preference Testing (HAPT)

Traditional adversarial testing relies on LLM-generated synthetic examples, which risk BDC due to overlap with pre-training corpora (Carlini et al., 2023). HAPT circumvents this by sourcing adversarial examples directly from humans, leveraging their ability to craft culturally nuanced and linguistically diverse claims that LLMs have not encountered during pre-training. This approach aligns with cognitive science findings that human-generated misinformation exhibits higher variability than synthetic text (Pennycook and Rand, 2021).

In this component, which utilises an approach similar to Chiang et al. (2024) for collecting human evaluation results, participants may be interested in certain events or news claims, so they were asked to submit both real and fake news stories to the HAPT system. Participants receive minimal guidelines (e.g., "*Write a plausible fake headline about climate change*") to encourage organic creativity. Submissions are filtered for redundancy and offensiveness, yielding a corpus of claims. For instance, a participant might submit the fabricated claim: "*The European Union has banned solar panels due to cancer risks from electromagnetic radiation.*" Each claim is classified by the target LLM (e.g., GPT-4) using zero-shot prompting. Following this process, we use the F-1 accuracy of its results for the output O_{HAPT}.

3.2 Component 2: Real-Time Web Agent with Asynchronous Validation (RTW-AV)

Static benchmarks like LIAR (Wang, 2017) suffer from temporal contamination, as their test sets often appear in LLM pre-training data (Razeghi et al., 2022). RTW-AV addresses this by evaluating LLMs on claims collected in real time from Internet, ensuring temporal novelty. By deferring ground-truth labeling until fact-checking consensus emerges (e.g., via Snopes[3] or PolitiFact[4]), the

[3] https://www.snopes.com/fact-check/
[4] https://www.politifact.com/factchecks/

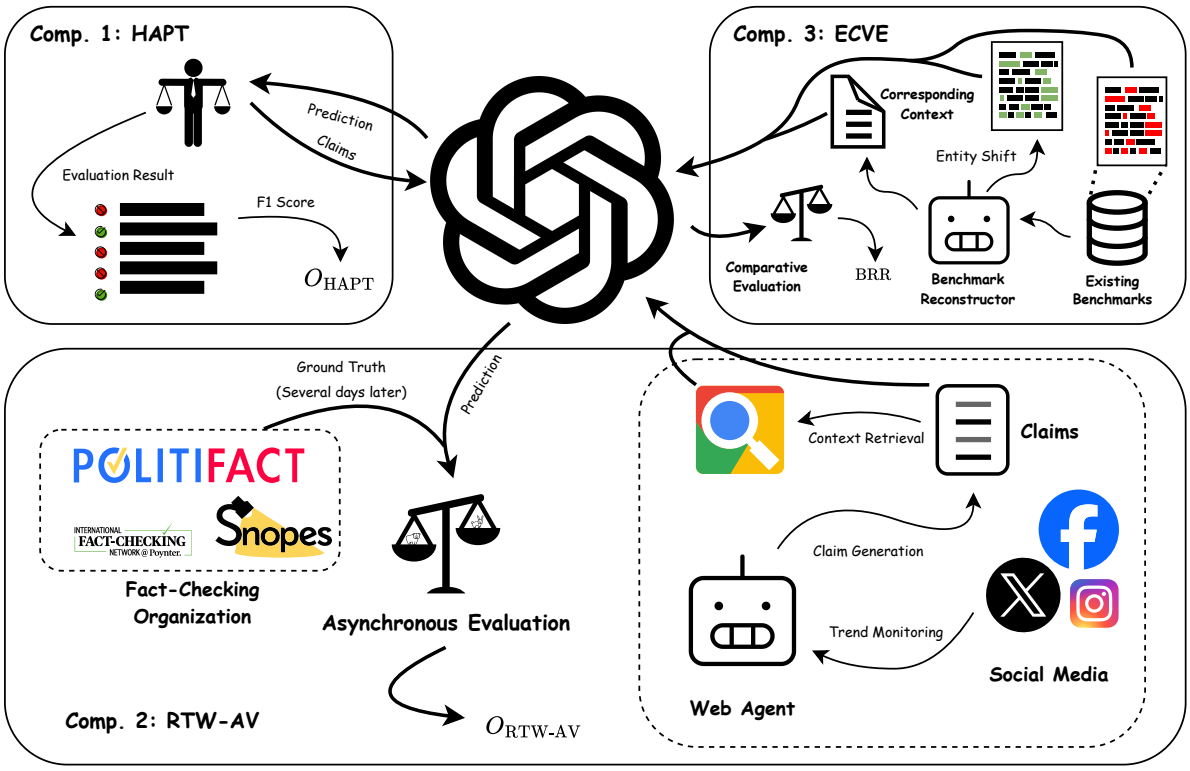

Figure 1: TripleFact Evaluation Framework

framework mirrors real-world misinformation verification workflows (Guo et al., 2022).

In practice, an agent can be employed to collect data, e.g., from a fixed source and transform them into news claims, and subsequently cause the evaluated LLMs to perform Retrieval-Augmented Detection based on each of the claims (Khan et al., 2022; Zeng and Gao, 2024). Finally, an asynchronous evaluation is performed, i.e., claims are labeled as "Real" or "Fake" several days post-evaluation using consensus from fact-checking platforms.

We introduced two metrics for the presentation of results in the RTW-AV, namely Time-Delayed Accuracy (TDA) and Context Utilization Score (CUS), which are defined below:

$$\text{TDA} = \frac{\text{Number of CCPV}}{\text{Total Claims}} \quad (1)$$

$$\text{CUS} = \frac{\text{CC} \geq 3 \text{ Retrieved Sources}}{\text{Total Claims}} \quad (2)$$

Where CCPV denotes Correct Classifications Post-Validation, and CC denotes Classifications Citing.

TDA can be considered as the evaluation result and CUS can be regarded as the confidence of the result, so the final output $O_{\text{RTW-AV}}$ of RTW-AV is calculated as the CUS-weighted TDA as follows:

$$O_{\text{RTW-AV}} = \text{CUS} \times \text{TDA} \quad (3)$$

3.3 Component 3: Entity-Controlled Virtual Environment (ECVE)

LLMs often exploit memorized entity associations (e.g., "Donald Trump" → "Fake News") rather than contextual reasoning (Wang et al., 2023; Moslemi and Zouaq, 2024). ECVE eliminates this bias by replacing real-world entities with synthetic counterparts, forcing models to rely on claim semantics, characteristics and the context provided.

Specifically, High-risk entities (e.g., organizations, politicians) are identified in existing benchmarks using named entity recognition. A rule-based reconstructor substitutes them with fictional analogs (e.g., "Pfizer" → "VaxGen," "WHO" → "Global Health Alliance"). For example, modify the original claim "*Pfizer suppressed reports of vaccine side effects.*" to "*VaxGen suppressed reports of vaccine side effects.*" Modified claims with their corresponding context are generated by the benchmark reconstructor, preserving claim semantics while altering entities.

Furthermore, two metrics have been introduced in the ECVE: Entity-Independent Accuracy (EIA) and Bias Reduction Ratio (BRR). These metrics are defined as follows:

$$\text{EIA} = \frac{\text{Number of CCMC}}{\text{Total Modified Claims}} \quad (4)$$

$$\text{BRR} = \text{EIA} - \text{OAUC} \tag{5}$$

where CCMC denotes correct classifications in modified claims, OAUC denotes original accuracy on unmodified claims. A negative BRR indicates the LLM relies on entity memorization; a positive value suggests well generalization.

3.4 Unified Evaluation Protocol

For the presentation of the final results of the whole TripleFact framework, we use the TripleFact Score (TFS) to refer to it in Equation 6. The structure is an average of HAPT and RTW-AV then weighted using BRR. The formula takes into account a combination of human preference and real-time evaluations, and uses ECVE to correct the final score. The whole framework considers multiple aspects in a comprehensive manner without consuming excessive computational resources.

$$\text{TFS} = (1 + \text{BRR})\frac{(O_{\text{HAPT}} + O_{\text{RTW-AV}})}{2} \tag{6}$$

The design philosophy behind the TripleFact framework is that HAPT evaluates the ability of LLMs to defend themselves against high-quality fake news by the human attack, RTW-AV evaluates the ability of LLMs to determine false information when confronted with information that is beyond their own knowledge base through information integration and reasoning to determine false information. While the first two are evaluating LLMs' ability to detect false information, ECVE is a mechanism to penalize LLMs' own internal general knowledge bias that affects the fake news detection task, through which we can learn whether LLMs make different decisions when confronted with news with and without prior knowledge. For example, if the tested LLM does not have any prior knowledge about *COVID-19*, theoretically neither *COVID-19* nor the fictitious *Zeta-19* will affect the conclusions made by the LLM, because the name is only a pronoun. At the same time if the tested LLM has a slight generalized knowledge bias, then the LLM should be able to make the same judgment after we provide the context about the fictitious entity, and if the LLM still makes the opposite decision (ruling out the randomness of the LLMs), then we can assume that the LLM is subjected to a serious contamination of the domain and should be penalized. The cooperation of the three modules together ensures the robustness of the TripleFact evaluation framework.

Model	Parameters	Context Length (Input/Output)	Knowledge Cut-off
o3-mini-2025-01-31	-	200k/100k	10/2023
o1-preview-2024-09-12	-	200k/100k	10/2023
o1-mini-2024-09-12	-	128k/65k	10/2023
gpt-4o-2024-08-06	-	128k/16k	10/2023
gpt-4o-mini-2024-07-18	-	128k/16k	10/2023
gpt-4-turbo-2024-04-09	-	128k/4k	12/2023
gpt-3.5-turbo-0125	175B	16k/4k	09/2021
Llama-3.3-70B	70B	128k/2k	12/2023
Llama-3.2-3B	3B	128k/2k	12/2023
Llama-3.2-1B	1B	8k/2k	12/2023
Llama-3.1-405B	405B	128k/2k	12/2023
Llama-3.1-70B	70B	128k/2k	12/2023
Llama-3.1-8B	8B	128k/2k	12/2023
Llama-3-70B	70B	8k/2k	12/2023
Llama-3-8B	8B	8k/2k	03/2023
DeepSeek-R1	671B	128k/32k	-
DeepSeek-V3	671B	128k/8k	-

Table 2: Comparison of the LLMs selected for the Experiment 1. Since the current state-of-the-art GPT family models are commercially closed-sourced, it is unavailable to obtain information about their parameters, while the DeepSeek family models do not disclose their knowledge cut-off.

4 Experiments

In this section, we conducted two experiments to validate the effectiveness of the TripleFact evaluation framework, Experiment 1: testing current state-of-the-art LLMs with the TripleFact framework to evaluate their fake news detection capabilities, and Experiment 2: injecting contamination into the tested LLMs to examine whether the TripleFact framework is effective and robust when the LLMs have been contaminated.

4.1 Experimental Settings

In Experiment 1, since our proposed TripleFact framework is lightweight and does not require access to the model's internal parameters and logits, we selected 17 models from 3 LLMs series for evaluation in order to make the experiment as representative as possible: the GPT series (Brown et al., 2020; OpenAI, 2024), the LLaMA series (Touvron et al., 2023a,b; Grattafiori et al., 2024), and the DeepSeek series (DeepSeek-AI et al., 2024, 2025), which are detailed in the Table 2.

For Experiment 2, we chose `gpt-4o-mini` for the contamination injection experiments, specifically we will fine-tune the contaminated corpus on the model with different degrees of contamination, and then test the metrics of the TripleFact framework to study the effectiveness and robustness. All experiments were carried out using the

Model \ Stage	HAPT	RTW-AV			ECVE		TFS
	O_{HAPT}	TDA	CUS	$O_{\text{RTW-AV}}$	EIA	BRR	
o3-mini-2025-01-31	0.608	0.721	0.992	0.715	0.272	0.013	0.670
- with medium	0.648	0.757	0.987	0.747	0.285	0.005	0.701
- with high	0.688	0.823	**1.0**	0.823	0.312	-0.011	0.747
o1-preview-2024-09-127	**0.718**	**0.878**	**1.0**	**0.878**	0.327	-0.012	**0.788**
o1-mini-2024-09-12	0.405	0.712	0.923	0.657	0.315	-0.042	0.509
gpt-4o-2024-08-06	0.575	0.743	0.986	0.733	0.339	0.030	0.674
gpt-4o-mini-2024-07-18	0.398	0.664	0.936	0.622	0.278	**0.031**	0.526
gpt-4-turbo-2024-04-09	0.413	0.687	0.841	0.577	0.303	-0.038	0.476
gpt-3.5-turbo-0125	0.512	0.582	0.858	0.499	0.315	0.021	0.516
Llama-3.3-70B	0.477	0.633	0.964	0.610	**0.342**	-0.034	0.525
Llama-3.2-3B	0.305	0.359	0.661	0.237	0.210	0.013	0.275
Llama-3.2-1B	0.317	0.398	0.623	0.248	0.136	-0.087	0.258
Llama-3.1-405B	0.558	0.738	0.967	0.714	0.341	-0.021	0.623
Llama-3.1-70B	0.465	0.639	0.892	0.570	0.318	-0.025	0.505
Llama-3.1-8B	0.332	0.562	0.869	0.488	0.295	0.015	0.416
Llama-3-70B	0.455	0.610	0.886	0.540	0.312	-0.098	0.449
Llama-3-8B	0.328	0.553	0.875	0.484	0.306	-0.018	0.399
DeepSeek-R1	0.652	0.803	**1.0**	0.803	0.332	-0.021	0.712
DeepSeek-V3	0.612	0.741	0.967	0.717	0.338	0.016	0.675

Table 3: Performance of state-of-the-art LLMs evaluated using the TripleFact framework across HAPT, RTW-AV, and ECVE components. Results highlight performance disparities across GPT, Llama, and DeepSeek model families, with top performers in each category **bolded** (highest) and underlined (second-highest). Notably, the o3-mini model has three reasoning modes (reasoning_effort), low, medium, and high, with low being the default.

OpenAI API[5], and the resource cost is shown in the Appendix A.1.

4.2 Exp. 1: Evaluating LLMs on Fake News Detection Tasks Using TripleFact

In this experiment, the core purpose is to use the TripleFact framework to evaluate the current state-of-the-art LLMs, and also to demonstrate that the framework is capable of benchmarking the fake news detection performance of LLMs.

In HAPT, ideally it should be crowd-testing by humans, and to simulate this procedure we referenced PolitiFact's recently validated fact-checking content and manually created 100 real and 100 fake news stories, respectively, modeled after their style to serve as test cases. Furthermore, we performed additional quality control procedure on the manually created content, as detailed in the Appendix A.2 for creation guidelines, examples, and statistical information.

In the RTW-AV, we use X API[6] to collect real-time news topics under X[7], 100 original tweets from *Top* and *Latest* for each topic, and then use the `gpt-4o-mini` model to generate claims for the captured tweets. The Google API is employed to obtain the top 10 pages (filtered according to the website use policy) searching for the topic as retrieved context information, which is sent into the LLM to be tested along with the claims for evaluation. The process is performed in real-time, so we set the evaluation to be conducted every 6 hours for a total evaluation period of 7 days. The details and examples are in Appendix A.3.

In ECVE, we tested with the LIAR2 dataset (Xu and Kechadi, 2024), an enhanced and expanded version of the LIAR dataset (Wang, 2017), which is the most representative benchmark in the field of fake news detection, with bug fixes and additional data up to August 2023, compared to the original LIAR dataset. The details and examples of this experiments are in Appendix A.4.

The results of Experiment 1, as summarized in Table 3, reveal critical insights into the performance of state-of-the-art LLMs under the TripleFact framework and highlight the framework's efficacy in benchmarking fake news detection capabilities. Overall, the o1-preview model achieved the highest TripleFact Score (TFS) of 0.788, outperforming other models across nearly all components. This model demonstrated exceptional robustness in human-adversarial testing (O_{HAPT}: 0.718) and real-time validation ($O_{\text{RTW-AV}}$: 0.878), suggesting superior generalization to novel, dynamically evolving misinformation. However, its relatively low BRR (-0.012) in the ECVE component indicates residual reliance on pre-trained entity associations, a limitation shared by most models. By contrast, GPT-4o achieved the second highest BRR (0.030) and EIA (0.339), showcasing stronger independence from entity memorization, though its TFS (0.674) lagged behind the top performer due to weaker HAPT performance. These results underscore the trade-offs between adversarial robustness, real-time adaptability, and entity bias mitigation in LLM-driven fake news detection.

A closer examination of individual components reveals systematic patterns. In the HAPT module, models exhibited significant variability, with O_{HAPT} ranging from 0.305 (Llama-3.2-3B) to 0.718 (o1-preview). The high O_{HAPT} of models like o3-mini-high (0.688) and DeepSeek-R1 (0.652) suggest that larger, more recent architectures better resist human-crafted deception. Conversely, the poor performance of smaller Llama variants (e.g., Llama-3.2-1B: 0.317) highlights the computational demands of adversarial robustness. In the RTW-AV component, nearly all models achieved near-perfect CUS (context utilization

[5] https://openai.com/api/
[6] https://docs.x.com/x-api/
[7] https://x.com/

scores $\geq$0.8), indicating effective use of retrieved evidence. However, TDA (time-delayed accuracy) varied widely, from 0.359 (Llama-3.2-3B) to 0.878 (o1-preview), reflecting disparities in temporal generalization. The ECVE results further exposed vulnerabilities: while GPT-4o-mini achieved the highest BRR (0.031), its low EIA (0.278) suggests that reducing entity bias may come at the cost of detection accuracy. This tension underscores the need for balanced evaluation frameworks that penalize over-reliance on memorized patterns without sacrificing performance.

The comparative analysis of model families yields additional insights. GPT-4o variants consistently outperformed Llama-3 series models in TFS, with the exception of Llama-3.1-405B (TFS: 0.623), which rivaled mid-tier GPT models. This suggests that scaling model parameters alone does not guarantee contamination resistance, as even the largest Llama variant (405B parameters) underperformed GPT-4o (TFS: 0.674). The DeepSeek models, particularly DeepSeek-R1 (TFS: 0.712), demonstrated competitive performance, likely due to specialized pre-training on fact-checking corpora. However, their negative BRR values (-0.021 for DeepSeek-R1) indicate persistent entity bias, a limitation shared by most non-GPT models. Notably, smaller models (e.g., Llama-3-8B: TFS 0.399) struggled across all components, emphasizing the computational-resource barriers to effective fake news detection. These findings validate the TripleFact framework's utility in disentangling model capabilities while exposing critical gaps in current LLM architectures.

4.3 Exp. 2: Contamination Injection into the TripleFact Framework

To evaluate the robustness of the TripleFact framework against BDC, we conducted a controlled contamination injection study using the `gpt-4o-mini` model. Building on the Experiment 1, we fine-tuned the model with varying contamination intensity (10%, 30%, 50%, 100%) across four BDC levels defined by Xu et al. (2024): *semantic*, *information*, *data*, and *label* level contamination. This experiment aims to quantify how different contamination types and degrees distort performance metrics in the TripleFact framework, thereby testing its resilience to compromised training data.

In the HAPT, it is the statement-related content that we use as a source of contamination. In the RTW-AV, the original tweets used to create the

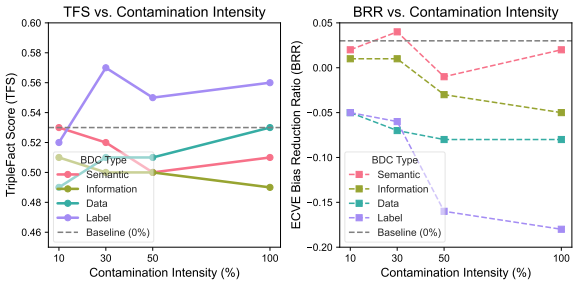

Figure 2: (**Left**) TFS trends show label level contamination artificially inflates scores, while semantic and information level contamination cause minor fluctuations. (**Right**) ECVE's BRR reveals severe bias under data and label level contamination, while the composite TFS maintains stability well enough to objectively evaluate the true capabilities of the model.

statements were included as contaminants in the fine-tuned corpus. In the ECVE, the LIAR2 dataset was modified to inject contamination at specified levels and intensity. For example, label level contamination involved appending test set samples (with labels) to the training corpus, while semantic level BDC introduced thematically similar but non-overlapping articles. The fine-tuning details are presented in the Appendix A.5.

The results in Table 4 and Figure 2 reveal nuanced effects of contamination across TripleFact components. The TripleFact framework demonstrates remarkable stability in evaluating fake news detection capabilities despite deliberate contamination, as evidenced by the marginal fluctuations in the TripleFact Score (TFS) across contamination levels and types. While contamination inflates metrics in the HAPT and RTW-AV components—e.g., label level contamination increased O_{HAPT} by 12% (0.40 $\rightarrow$ 0.52) and $O_{\text{RTW-AV}}$ by 23% (0.62 $\rightarrow$ 0.85)—the ECVE component systematically penalizes these gains. The BRR of ECVE plummeted to -0.18 under full label level contamination, offsetting inflated performance elsewhere. This indicates that the TripleFact framework does not merely measure raw classification accuracy but instead evaluates a model's ability to generalize beyond memorized patterns. For instance, while label level BDC models achieved higher $O_{\text{RTW-AV}}$ (0.85), their negative BRR values reveal overreliance on pre-trained entity associations, which ECVE isolates and penalizes. Thus, the stability of TFS (0.49–0.57 across all contamination levels, i.e. $\leq$4% absolute error) reflects the framework's capacity to disentangle genuine detection skill from

TripleFact \ BDC Level	Baseline	Semantic				Information				Data				Label				Avg.
	0%	10%	30%	50%	100%	10%	30%	50%	100%	10%	30%	50%	100%	10%	30%	50%	100%	
HAPT	0.40	0.40	0.39	0.42	0.42	0.39	0.41	0.48	0.47	0.42	0.44	0.45	0.47	0.36	0.43	0.48	0.52	0.46
RTW-AV	0.62	0.63	0.60	0.58	0.57	0.61	0.57	0.54	0.56	0.62	0.61	0.65	0.68	0.73	0.78	0.82	0.85	0.69
ECVE	0.03	0.02	0.04	-0.01	0.02	0.01	0.01	-0.03	-0.05	-0.05	-0.07	-0.08	-0.08	-0.05	-0.06	-0.16	-0.18	-0.05
TFS	0.53	0.53	0.52	0.50	0.51	0.51	0.50	0.50	0.49	0.49	0.51	0.51	0.53	0.52	0.57	0.55	0.56	0.52

Table 4: TripleFact evaluation results under contamination injection, values are the final output for each component.

contamination artifacts.

5 Conclusion

The rise of LLMs has fundamentally altered the landscape of fake news detection task, rendering traditional evaluation paradigms obsolete due to BDC and memorization risks. This paper highlights the critical limitations of static benchmarks in the LLM era, where models' exposure to vast pretraining corpora conflates genuine reasoning with pattern memorization. By proposing the TripleFact framework, we address these challenges through a contamination-aware methodology that prioritizes temporal robustness, adversarial generalization, and entity-agnostic evaluation. Experiments across 17 LLMs demonstrate that TripleFact effectively isolates detection capabilities from contamination artifacts, with its composite TripleFact Score remaining stable even under deliberate contamination injection. Future work should extend TripleFact to multilingual and multimodal misinformation scenarios while addressing its computational and human-in-the-loop dependencies.

Limitations

While TripleFact advances contamination-resistant evaluation, several limitations warrant consideration. HAPT requires crowd-sourced adversarial examples, which may lack scalability and consistency across cultural contexts (Li et al., 2024a). Synthetic entity generation of ECVE may alter claim semantics, particularly for niche domains (e.g., medical misinformation) (Waszak et al., 2018). The framework is validated primarily on English-language models; performance in low-resource languages remains untested (Yan and Xu, 2024). RTW-AV's deferred ground-truth validation assumes consensus among fact-checking platforms, which may lag for emerging claims. Nevertheless, this experiment establishes TripleFact as a versatile, contamination-resistant framework for benchmarking LLMs in fake news detection task.

Ethical Considerations

All datasets, models, and checkpoints used in this study strictly adhere to their respective use policies. The LIAR2 dataset, which forms the basis of our experiments, consists of publicly available political statements and does not contain sensitive or personally identifiable information beyond the names of public figures, thereby minimizing privacy risks. Furthermore, no harmful, hateful, or impermissible content was generated, stored, or disseminated during data processing, model training, or inference. Notably, while the data from X and PolitiFact are permissible for personal research use under their licenses, their redistribution is explicitly prohibited. Consequently, these datasets cannot be publicly released alongside this work. But in fact, our decision to incorporate these sources stems from their status as the most up-to-date data for mitigating data contamination risks, which escalate significantly as datasets age and proliferate across public repositories. To ensure reproducibility while aligning with ethical guidelines, we recommend that researchers adopting the TripleFact framework similarly prioritize contemporaneous data for evaluation rather than replicating existing benchmarks. Importantly, the framework itself is agnostic to the public availability of input data, as its design does not depend on proprietary or restricted sources. Finally, we clarify that PolitiFact's data was not directly utilized in our experiments but served as a supplementary reference; its role can be fully substituted with the open-source LIAR/LIAR2[8] datasets without compromising methodological validity. All processes were conducted in compliance with institutional and domain-specific ethical standards, with no foreseeable risks to individuals or communities.

Acknowledgments

We acknowledge the support from OpenAI Inc. for this work.

[8] https://huggingface.co/datasets/chengxuphd/liar2

References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, and et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, and et al. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Chunyuan Deng, Yilun Zhao, Yuzhao Heng, Yitong Li, Jiannan Cao, Xiangru Tang, and Arman Cohan. 2024. Unveiling the spectrum of data contamination in language model: A survey from detection to remediation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16078–16092, Bangkok, Thailand. Association for Computational Linguistics.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2023. Investigating data contamination in modern benchmarks for large language models. *Preprint*, arXiv:2311.09783.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12039–12050, Bangkok, Thailand. Association for Computational Linguistics.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, and et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Benjamin D. Horne, Jeppe Nørregaard, and Sibel Adali. 2019. Robust fake news detection over time and attack. *ACM Trans. Intell. Syst. Technol.*, 11(1).

Daphne Ippolito, Florian Tramer, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher Choquette Choo, and Nicholas Carlini. 2023. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 28–53, Prague, Czechia. Association for Computational Linguistics.

Minhao Jiang, Ken Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. 2024. Does data contamination make a difference? insights from intentionally contamination pre-training data for language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Kashif Khan, Ruizhe Wang, and Pascal Poupart. 2022. WatClaimCheck: A new dataset for claim entailment and inference. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1293–1304,

Dublin, Ireland. Association for Computational Linguistics.

Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pages 1–49. Citeseer.

Stephan Lewandowsky. 2025. Free speech, fact checking, and the right to accurate information. *Science*, 387(6734).

Changmao Li and Jeffrey Flanigan. 2024. Task contamination: Language models may not be few-shot anymore. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18471–18480.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 84799–84838. Curran Associates, Inc.

Sihang Li, Jin Huang, Jiaxi Zhuang, Yaorui Shi, Xiaochen Cai, Mingjun Xu, Xiang Wang, Linfeng Zhang, Guolin Ke, and Hengxing Cai. 2025. Scilitllm: How to adapt llms for scientific literature understanding. In *The Thirteenth International Conference on Learning Representations*.

Xiang Li, Yunshi Lan, and Chao Yang. 2024b. Treeeval: Benchmark-free evaluation of large language models through tree planning. *Preprint*, arXiv:2402.13125.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2024c. Latesteval: Addressing data contamination in language model evaluation through dynamic and time-sensitive test construction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18600–18607.

Yiheng Liu, Hao He, Tianle Han, Xu Zhang, Mengyuan Liu, Jiaming Tian, Yutong Zhang, Jiaqi Wang, Xiaohui Gao, Tianyang Zhong, Yi Pan, Shaochen Xu, Zihao Wu, Zhengliang Liu, Xin Zhang, Shu Zhang, Xintao Hu, Tuo Zhang, Ning Qiang, and 2 others. 2024. Understanding llms: A comprehensive overview from training to inference. *Preprint*, arXiv:2401.02038.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

Yougang Lyu, Lingyong Yan, Shuaiqiang Wang, Haibo Shi, Dawei Yin, Pengjie Ren, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. 2024. KnowTuning: Knowledge-aware fine-tuning for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14535–14556, Miami, Florida, USA. Association for Computational Linguistics.

Mehrnaz Moslemi and Amal Zouaq. 2024. TagDebias: Entity and concept tagging for social bias mitigation in pretrained language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1553–1567, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. Fine-tuning or retrieval? comparing knowledge injection in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 237–250, Miami, Florida, USA. Association for Computational Linguistics.

Eleftheria Papageorgiou, Christos Chronis, Iraklis Varlamis, and Yassine Himeur. 2024. A survey on the use of large language models (llms) in fake news. *Future Internet*, 16(8).

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.

Gordon Pennycook and David G Rand. 2021. The psychology of fake news. *Trends in cognitive sciences*, 25(5):388–402.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Y.M. Rocha, G.A. de Moura, G.A. Desidério, C.H. de Oliveira, F.D. Lourenço, and L.D. de Figueiredo Nicolete. 2021. The impact of fake news on social media and its influence on health during the covid-19 pandemic: A systematic review. *Journal of Public Health*, pages 1–10.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

Nataliya Stepanova and Björn Ross. 2023. Temporal generalizability in multimodal misinformation detection. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 76–88, Singapore. Association for Computational Linguistics.

Yifan Sun, Han Wang, Dongbai Li, Gang Wang, and Huan Zhang. 2025. The emperor's new clothes in benchmarking? a rigorous examination of mitigation strategies for llm benchmark data contamination. *Preprint*, arXiv:2503.16402.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, and et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *Preprint*, arXiv:2304.06588.

Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. 2023. A causal view of entity bias in (large) language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15173–15184, Singapore. Association for Computational Linguistics.

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

Przemyslaw M. Waszak, Wioleta Kasprzycka-Waszak, and Alicja Kubanek. 2018. The spread of medical fake news in social media – the pilot quantitative study. *Health Policy and Technology*, 7(2):115–118.

Bingbing Wen, Chenjun Xu, Bin HAN, Robert Wolfe, Lucy Lu Wang, and Bill Howe. 2024. From human to model overconfidence: Evaluating confidence dynamics in large language models. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 3367–3378, New York, NY, USA. Association for Computing Machinery.

Chunqiu Steven Xia, Yinlin Deng, and Lingming Zhang. 2024. Top leaderboard ranking = top coding proficiency, always? evoeval: Evolving coding benchmarks via llm. *Preprint*, arXiv:2403.19114.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. 2024. Benchmark data contamination of large language models: A survey. *Preprint*, arXiv:2406.04244.

Cheng Xu and M-Tahar Kechadi. 2023. Fuzzy deep hybrid network for fake news detection. In *Proceedings of the 12th International Symposium on Information and Communication Technology*, SOICT '23, page 118–125, New York, NY, USA. Association for Computing Machinery.

Cheng Xu and M-Tahar Kechadi. 2024. An enhanced fake news detection system with fuzzy deep learning. *IEEE Access*, 12:88006–88021.

Cheng Xu, Jing Wang, Tianlong Zheng, Yue Cao, and Fan Ye. 2022. Prediction of prognosis and survival of patients with gastric cancer by a weighted improved random forest model: an application of machine learning in medicine. *Archives of Medical Science*, 18(5):1208–1220.

Cheng Xu and Nan Yan. 2023. AROT-COV23: A dataset of 500k original arabic tweets on COVID-19. In *4th Workshop on African Natural Language Processing*.

Nan Yan and Cheng Xu. 2024. Decolonizing african NLP: A survey on power dynamics and data colonialism in tech development. In *5th Workshop on African Natural Language Processing*.

Jiahao Ying, Yixin Cao, Bo Wang, Wei Tang, Yizhe Yang, and Shuicheng Yan. 2024. Have seen me before? automating dataset updates towards reliable and timely evaluation. *Preprint*, arXiv:2402.11894.

Zhuohao Yu, Chang Gao, Wenjin Yao, Yidong Wang, Zhengran Zeng, Wei Ye, Jindong Wang, Yue Zhang, and Shikun Zhang. 2024. Freeeval: A modular framework for trustworthy and efficient evaluation of large language models. *Preprint*, arXiv:2404.06003.

Reza Zafarani, Xinyi Zhou, Kai Shu, and Huan Liu. 2019. Fake news research: Theories, detection strategies, and open problems. In *Proc of the 25th ACM SIGKDD Int'l. Conf on Knowledge Discovery & Data Mining*, KDD'19, page 3207–3208, New York, NY, USA. ACM.

Fengzhu Zeng and Wei Gao. 2024. JustiLM: Few-shot justification generation for explainable fact-checking of real-world claims. *Transactions of the Association for Computational Linguistics*, 12:334–354.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5).

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. 2024a. Dyval: Graph-informed dynamic evaluation of large language models. In *The Twelfth International Conference on Learning Representations*.

Kaijie Zhu, Jindong Wang, Qinlin Zhao, Ruochen Xu, and Xing Xie. 2024b. Dyval 2: Dynamic evaluation of large language models by meta probing agents. *Preprint*, arXiv:2402.14865.

A Experiment Details

In all experiments, all experiments were run three times and then averaged as a result. For the models that were able to specify the temperature, we set the `temperature=0`. The OpenAI model calls are all from the OpenAI API. The other open source models use versions from HuggingFace[9], and they come from Meta Llama[10] and DeepSeek AI[11], the experiments were conducted on the servers equipped with NVIDIA H100 GPUs using the FastChat[12] (Zheng et al., 2023) framework.

A.1 Resources Cost

The OpenAI credits consumed in experiments are shown in Table 5, and the consumption time of the open source model is demonstrated in Table 6, where since the RTW-AV experiment is not evaluated at once so its time is inferred from the time of a single evaluation.

A.2 HAPT Evaluation

In the HAPT experiment, we first referenced PolitiFact's most recent fact-checking content (e.g., style and presentation), and then manually created 100 real and 100 fake news stories for the evaluation, all of which centered on the creation of extremely realistic news (whether real or fake). Specifically, for the real news, make sure the content and style are correct is all that is needed; and for the fake news, the instruction here is not just to create fake news, but to create fake news that "fools" LLMs. For example, a fake news story that we created: "*In 2024, Russia won the Russo-Ukrainian War due to its great superiority and signed the Kiev Liberation Treaty with Ukraine, which ended the war by ceding the Crimean Peninsula to Russia as a price.*"

All manually generated news was produced by one author and subsequently reviewed by another, and was only included in the evaluation case when all authors agreed on the truthfulness and style of the news. Specifically, we generate a total of 116 true news and 102 fake news, and after discarding the news that did not pass the double-check, we get 100 true and 100 fake news, respectively.

The test prompts used for this evaluation are exemplified by the GPT series:

```
<|system|>
Classify the given political
    statement with two label: ""false
```

Model Name	HAPT	RTW-AV	ECVE
o3-mini-2025-01-31	0.63	9.91	14.12
- with medium	1.98	30.81	43.24
- with high	4.19	59.98	87.21
o1-preview-2024-09-12	8.94	277.82	407.22
o1-mini-2024-09-12	0.34	4.25	5.26
gpt-4o-2024-08-06	0.07	2.98	4.92
gpt-4o-mini-2024-07-18	<0.01	0.35	0.48
gpt-4-turbo-2024-04-09	0.26	10.32	19.67
gpt-3.5-turbo-0125	0.01	0.54	0.74

Table 5: OpenAI API cost of a single experiment by TripleFact framework ($).

Model Name	HAPT	RTW-AV	ECVE
Llama-3.3-70B	<0.5	<0.5	<0.5
Llama-3.2-3B	<0.5	<0.5	<0.5
Llama-3.2-1B	<0.5	<0.5	<0.5
Llama-3.1-405B	<0.5	~0.5	~1
Llama-3.1-70B	<0.5	<0.5	<0.5
Llama-3.1-8B	<0.5	<0.5	<0.5
Llama-3-70B	<0.5	<0.5	<0.5
Llama-3-8B	<0.5	<0.5	<0.5
DeepSeek-R1	<0.5	~1	~1.5
DeepSeek-V3	<0.5	~1	~1.5

Table 6: GPU cost of a single experiment by TripleFact framework (hour).

[9] https://huggingface.co/
[10] https://huggingface.co/meta-llama
[11] https://huggingface.co/deepseek-ai
[12] https://github.com/lm-sys/FastChat

```
      "" or ""true""\n
3  </s>
4  <|user|>
5  Only provide a JSON object with the
       keys \"label\" ""false"" or ""
       true"" based on the truthfulness
       of the statement.\n "false" means
        it is a fake news, and "true"
       means it is a real news.\n If you
        don't know, choose a label based
        on your reasoning.\n Statement:
       {statement}
6  </s>
```

A.3 RTW-AV Evaluation

In the RTW-AV experiment, we use X API to capture tweets every 6 hours for a total of 7 days, and then make `gpt-4o-mini` to generate claims based on the captured tweets, using the following prompt:

```
1  <|system|>
2  You need help generating news
       statements based on tweets you
       received.\n
3  </s>
4  <|user|>
5  Only provide a JSON object with the
       keys \"statements\" based on the
       tweets you received.\n If not
       please return the empty string.
       If there is more than one
       statement, use "\n" to split them
       .
6  </s>
```

Subsequently, Google API[13] was used to search all the statements one by one, crawling them in the top 10 pages retrieved (filtered according to the website use policy). Following this process, we collected a total of 157 statements and the corresponding 1,570 web pages, each of which was manually verified to ensure that they were valid. And then sent the statement one by one with the corresponding retrieved page to the LLM under test using the following prompt to cause LLM to generate a truthfulness judgment of the claim:

```
1  <|system|>
2  Classify the given political
       statement with two label: ""false
       "" or ""true""\n You can retrieve
        information from the fetched
       webpage to help you make
       decisions, but the information
       you use must be labeled with [
       webpage-number] to indicate which
        webpage you are using in your
       justification.\n
3  </s>
4  <|user|>
5  Only provide a JSON object with the
       keys \"label\" ""false"" or ""
       true"" based on the truthfulness
       of the statement, and provide the
        corresponding justification
       using the key \"justification\".\
       n "false" means it is a fake news
       , and "true" means it is a real
       news.\n If you don't know, choose
        a label based on your reasoning
       .\n Statement: {statement}\n
       Webpage: {webpage}
6  </s>
```

[13] https://developers.google.com/custom-search/v1/overview

A.4 ECVE Evaluation

For the ECVE experiments, we still employed `gpt-4o-mini` to help us with the entity shift step. Specifically, for our experiment dataset (the LIAR2 test set), we use the following prompt to let the model generate statements and contexts after entity shift:

```
1  <|system|>
2  Generate an entity-shifted claim and
       corresponding context by
       replacing real-world entities
       with fictional analogs while
       preserving claim semantics. Do
       not mention in context that the
       entities in it are fictional.\n
3  </s>
4  <|user|>
5  Return only the JSON object with the
       keys \"statement_revised\" and \"
       context\". Ensure fictional names
        are completely original and not
       similar to real entities.\n Do
       not mention the existence of
       fictions in the generated \"
       statement_revised\" and \"context
       \", especially do not mention in
       \"context\" that the entities are
        fictional.\n For \"
       statement_revised\", claim with
       real entities replaced by
       fictional counterparts.\n
       Identify named entities (people,
       organizations, geopolitical
       entities)\n"Replace each entity
       with a unique fictional name\n
       Maintain original grammatical
       structure and claim meaning\n\n
       For \"context\", brief
       explanatory text for fictional
       entities\n One sentence per
       fictional entity, use format '[
       Name] is a [description]'\n Keep
       descriptions generic (e.g., 'an
       American politician', 'a
       pharmaceutical company')\n\n
       Example 1:\n 'Statement': 'Trump
       says we should protect the moat
       of AI in the US.'\n '
       statement_revised': 'Wannetta
       says we should protect the moat
       of AI in the US.'\n 'context': '
       Wannetta is an American
       politician.'\n Example 2:\n '
       Statement': 'Pfizer suppressed
       reports of vaccine side effects
       .'\n 'statement_revised': 'VaxGen
        suppressed reports of vaccine
```

```
        side effects.'\n 'context': '
        VaxGen is a pharmaceutical
        company.'\n Example 3:\n '
        Statement': 'WHO releases 2025
        update to the International
        Classification of Diseases.'\n '
        statement_revised': 'Global Human
         Health Institute (GHHI) releases
         2025 update to the International
         Classification of Diseases.'\n '
        context': 'Global Human Health
        Institute (GHHI) is a global
        health organization.'\n\n
        Statement: {statement}
</s>
```

Here is a randomly selected data after an entity shift step:

```
<|label|>
1
</s>
<|revised statement|>
Three healers from the same medical
    center 'die suddenly' in the same
     week after the medical center
    mandated a fourth Zeta-19 vaccine
     for employees.
</s>
<|context|>
Healers are medical professionals
    specializing in patient care. The
     medical center is a healthcare
    facility providing various
    medical services. Zeta-19 is a
    vaccine developed to combat a
    specific viral infection.
</s>
<|original statement|>
Three doctors from the same hospital
    'die suddenly' in the same week
    ,"" after the hospital mandated a
     fourth COVID-19 vaccine for
    employees.
</s>
```

The main reason we employ `gpt-4o-mini` for entity shift operations in our main experiment is that it is cost-effective and more likely to become generalized. But also to understand the performance of the current state-of-the-art models on this work, we performed an additional comparison, i.e., we used the `o3-mini` model to perform the same operation and then manually checked all the entity-shifted data, the statistical results and the costs are provided in the Table 7. With this additional experiment, we found that `o3-mini` is essentially perfect for this task, especially with `reasoning_effort=high`. And although `gpt-4o-mini` consumes very few resources, it still has a gap with `o3-mini` in terms of reliability (~15%). Therefore, we suggest that if a more general and economical LLM such as `gpt-4o-mini` is used for entity shift operations, it is best to require a more careful manual review by the evaluator to ensure the reliability of the LLM generated content.

Model Name	Cost ($)	Reliability (%)
gpt-4o-mini-2024-07-18	0.21	85.10
o3-mini-2025-01-31	6.40	99.61
- with medium	20.70	99.96
- with high	44.65	100.0

Table 7: OpenAI API cost and reliability of the entity shift procedure. Reliability refers to the percentage of entity-shifted data generated by LLM that can be directly adopted without any modification.

A.5 Fine-tuning for Contamination Injection Experiments

In the HAPT, extensive news data, e.g., GDELT[14] (Leetaru and Schrodt, 2013), and task information related to fake news detection (e.g., Wikipedia[15]) are seen as a source of semantic level BDC. The information level BDC data used were obtained from searches conducted on low-confidence social platforms (e.g., X API) using the human generated claims. Data level and label level BDC, on the other hand, both use test data directly for fine-tuning, the difference being that the former does not include labels. In RTW-AV, a similar approach to HAPT is followed, but with the difference that the information level BDC data is derived from the retrieved web content and also the original tweets.

In the ECVE, for semantic level BDC experiments, we the same source of the HAPT part and clean it as necessary to serve as a fine-tuned corpus. For information level BDC experiments, we use Paper-with-code[16] to collect research papers using the benchmark under the LIAR and LIAR2 datasets, and incorporate them into the corpus after PDF parsing. The data level BDC experiments, on the other hand, are directly contaminated with data from the training (with labels) sets of LIAR2. Unlike the data level, we use test set (with labels) to the labeled level contaminated corpus.

Since the GPT series of models does not have an open interface for pre-training, we convert the prepared contaminated corpus into the form of Q&A pairs for injecting contamination, the basis is that a number of studies have also demonstrated that knowledge can be injected through supervised fine-tuning (SFT) (Ovadia et al., 2024; Lyu et al., 2024;

[14] https://www.gdeltproject.org/
[15] https://wikipedia.org/
[16] https://paperswithcode.com/dataset/liar

Li et al., 2025). The three components have the same percentage of total contamination data. Except for the contamination injection part, the rest of the fine-tuning corpus is randomly selected from FLAN (Longpre et al., 2023).

The `gpt-4o-mini` model was fine-tuned for 3 epochs per configuration (learning rate: 1×10^{-3}, batch size: 16). Post fine-tuning, each contaminated model was evaluated on the original LIAR2 test set using the TripleFact framework under identical conditions to Experiment 1. We conducted 16 total runs (4 BDC levels × 4 contamination intensity) and measured degradation in TripleFact Score (TFS), HAPT, RTW-AV, and ECVE metrics.