

# LiveFact: A Dynamic, Time-Aware Benchmark for LLM-Driven Fake News Detection

Cheng Xu<sup>1,4</sup> Changhong Jin<sup>1</sup> Yingjie Niu<sup>1</sup> Nan Yan<sup>2,4</sup>  
Yuke Mei<sup>4</sup> Shuhao Guan<sup>1</sup> Liming Chen<sup>3</sup> M-Tahar Kechadi<sup>1</sup>

<sup>1</sup> University College Dublin <sup>2</sup> Georgia Institute of Technology

<sup>3</sup> Dalian University of Technology <sup>4</sup> Bebx

cheng.xu1@ucdconnect.ie tahar.kechadi@ucd.ie

## Abstract

The rapid development of Large Language Models (LLMs) has transformed fake news detection and fact-checking tasks from simple classification and fact-checking to complex reasoning. However, evaluation frameworks have not kept pace. Current benchmarks are static, making them vulnerable to benchmark data contamination (BDC) and ineffective at assessing reasoning under temporal uncertainty. To address this, we introduce **LiveFact**<sup>1</sup> a continuously updated benchmark that simulates the real-world "fog of war" in misinformation detection. LiveFact uses dynamic, temporal evidence sets to evaluate models on their ability to reason with evolving, incomplete information rather than on memorized knowledge. We propose a dual-mode evaluation: Classification Mode for final verification and Inference Mode for evidence-based reasoning, along with a component to monitor BDC explicitly. Tests with 22 LLMs show that open-source Mixture-of-Experts models, such as Qwen3-235B-A22B, now match or outperform proprietary state-of-the-art systems. More importantly, our analysis finds a significant "reasoning gap." Capable models exhibit epistemic humility by recognizing unverifiable claims in early data slices—an aspect traditional static benchmarks overlook. LiveFact sets a sustainable standard for evaluating robust, temporally aware AI verification.

## 1 Introduction

Fake news detection has long been a foundational task in Natural Language Processing (NLP) and information verification, evolving from simple linguistic analysis to complex system verifications (Guo et al., 2022). Early research treated this as a binary classification problem, relying heavily on feature engineering, such as writing style analysis, word-frequency statistics, and sentiment patterns

<sup>1</sup><https://github.com/bebxy/livefact>

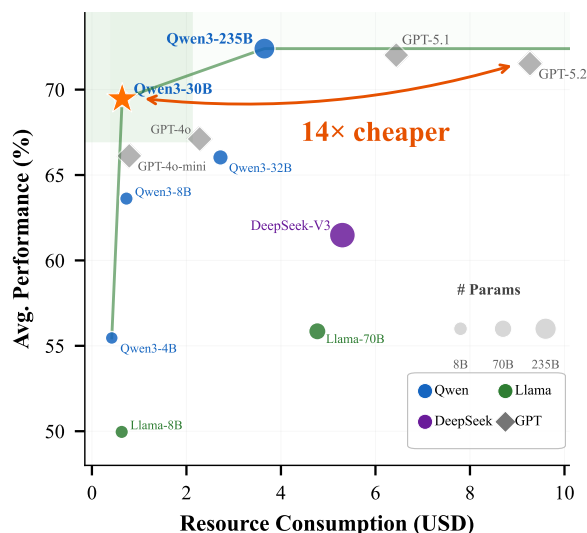


Figure 1: Cost-performance trade-off on LiveFact (Nov. 2025). Qwen3-235B-A22B-Instruct-2507 achieves the best performance (72.4%), while Qwen3-30B-A3B-Instruct-2507 provides optimal cost-efficiency at 14× lower cost than comparable GPT models.

(Zhou and Zafarani, 2020). As social media platforms proliferated, the scope expanded to include propagation analysis, where researchers modeled the spread of information across networks to identify anomalies (Rocha et al., 2021; Sadeghi et al., 2022; Mishra, 2020; Chen et al., 2023). These traditional approaches laid the groundwork for automated verification, but they were largely constrained by reliance on static features that cannot capture the semantic depth required for effective detection of fake news (Xu et al., 2025b; Xu and Kechadi, 2025).

The advent of Large Language Models (LLMs) has fundamentally transformed this landscape (OpenAI, 2024; Anthropic, 2024; Team et al., 2024), shifting the paradigm from pattern recognition to semantic reasoning (Zhang et al., 2024b; Havrilla et al., 2024; Tong et al., 2025). Modern LLMs possess unprecedented capabilities in knowl-

edge retrieval, context understanding, and logical inference, enabling them to approach fact-checking with a level of sophistication that mirrors human cognition (Zhang et al., 2025; Xiao et al., 2025; Ma et al., 2024b). The task has consequently evolved from simple "true/false" classification into a complex reasoning challenge. Systems are now required to verify claims against multi-hop evidence, identify contradictions in long-form context, and provide interpretable explanations for their verdicts (Liao et al., 2023; Atanasova et al., 2020; Alhindi et al., 2018). This shift has raised the ceiling for what is possible in automated fake news detection.

However, there is a critical misalignment between the rapid advancement of LLMs and the stagnation of the benchmarks used to evaluate them (Xu and Yan, 2025). The field of fake news detection has not kept pace with the generative AI revolution, leading to a "capability-evaluation gap." Current mainstream benchmarks remain static, offering fixed snapshots of information that cannot challenge the dynamic capabilities of modern models (e.g., OpenAI's latest GPT-5.2 model has a knowledge cutoff of August 2025<sup>2</sup>). They typically present a "God-view" scenario in which all relevant evidence is perfectly aggregated, ignoring the temporal nuances and information scarcity, which define real-world fake news detection. Furthermore, because these static datasets are frequently ingested into the massive training corpora of LLMs, they carry high risks of Benchmark Data Contamination (BDC) (Xu et al., 2024; Sun et al., 2025), making it difficult to discern whether a model is genuinely reasoning or merely reciting memorized training data (Lee et al., 2022; Sainz et al., 2023; Zhou et al., 2023; McIntosh et al., 2024; Deng et al., 2024).

The root cause of these limitations lies in the static nature of dataset construction, which directly conflicts with the evolving nature of news and the continuous pre-training of LLMs (White et al., 2025; Chen et al., 2025b). To bridge this gap, we must fundamentally rethink evaluation paradigms. A robust solution requires transitioning from static archives to dynamic, living benchmarks that continuously update with the news cycle. Furthermore, evaluation must move beyond simple fact retrieval to test reasoning under uncertainty, simulating the "fog of war" where evidence is incomplete or evolving. Only by controlling for time and BDC can we

accurately measure an LLM's true utility as a fake news detection assistant.

To address these issues, we introduce **LiveFact**, a benchmark designed to rigorously evaluate fake news detection with LLMs in a dynamic environment. LiveFact specifically contributes the following:

1. **Dynamic Benchmarking:** We continuously update the dataset on a monthly basis using the latest news, ensuring zero-shot evaluation on unseen events and minimizing BDC risk.
2. **Sustainable Contamination Monitoring:** We employ Semantic Sensitivity Amplifier (SSA) framework (Xu et al., 2025b) with entity shift mechanism to explicitly measure and mitigate memorization risks.
3. **Fine-Grained Temporal Evidence:** We utilize dynamic time-sliced evidence sets (e.g.,  $E^{(-3)}$  vs.  $E^{(+3)}$ ) to simulate real-world information evolution.
4. **Dual-Mode Evaluation:** We employ both Inference Mode (reasoning based on available evidence) and Classification Mode (knowledge-based verification) to disentangle reasoning skills from internal knowledge.

## 2 Related Work

The early era of fake news detection was defined by static datasets and feature-based learning. Pioneering benchmarks such as LIAR (Wang, 2017) and FakeNewsNet (Shu et al., 2020) provided labeled short statements from political contexts, enabling classifiers that rely on linguistic cues and metadata. Subsequent iterations, such as LIAR-PLUS (Alhindi et al., 2018) and LIAR2 (Xu and Kechadi, 2024, 2023), attempted to add evidence context, while other datasets focused on social context and propagation graphs (Meyers et al., 2020; Gong et al., 2023; Zhu et al., 2024; Si et al., 2023; Huang et al., 2025; Cao et al., 2025; Zheng et al., 2025). While foundational, these benchmarks are not adequate for LLMs. They are typically small-scale, domain-specific (often limited to U.S. politics), and, most critically, static. Thus, modern LLMs, which have likely seen this data during pre-training, can solve them via memorization rather than reasoning (White et al., 2025). Furthermore, the evidence in these older datasets is often simplistic, lacking the multi-hop complexity required

<sup>2</sup><https://platform.openai.com/docs/models/gpt-5.2>

Benchmark	Author	Type	Update	Evidence	Temporal	BDC Control	Evaluation Focus
LIAR	Wang (2017)	Static	✗	✗	✗	✗	Classification
LIAR-PLUS	Alhindi et al. (2018)	Static	✗	✓	✗	✗	Classification
FEVER	Thorne et al. (2018)	Static	✗	✓	✗	✗	Classification
HotpotQA	Yang et al. (2018)	Static	✗	✓	✗	✗	Reasoning
NELA-GT-2018	Nørregaard et al. (2019)	Static	✗	✗	✗	✗	Classification
FakeNewsNet	Shu et al. (2020)	Static	✗	✗	✗	✗	Classification
HoVer	Jiang et al. (2020)	Static	✗	✓	✗	✗	Reasoning
PolitiHop	Ostrowski et al. (2021)	Static	✗	✓	✗	✗	Reasoning
CHECKED	Yang et al. (2021)	Static	✗	✓	✓	✗	Classification
Weibo23	Liu et al. (2024)	Static	✗	✓	✗	✗	Classification
MCFEND	Li et al. (2024)	Static	✗	✓	✗	✗	Classification
FineFake	Zhou et al. (2024)	Static	✗	✓	✗	✗	Classification
MultiHoax	Shafiei et al. (2025)	Static	✗	✓	✗	✗	Reasoning
MMFakeBench	Liu et al. (2025)	Static	✗	✓	✗	✗	Classification
MPPFND	Zhao et al. (2025)	Static	✗	✓	✓	✗	Classification
AdvFake	Chen et al. (2025a)	Dynamic	✗	✓	✓	Medium	Classification
TripleFact	Xu and Yan (2025)	Dynamic	✗	✓	✗	Medium	Classification
LiveFact	Ours	Dynamic	✓	✓	✓	High	Reasoning & Classification

Table 1: Comparison of LiveFact with existing representative fake news detection benchmarks.

to evaluate current generative models (Ostrowski et al., 2021).

As LLMs have become widely used, research has shifted toward more complex, reasoning-intensive frameworks. MultiHoax (Shafiei et al., 2025) introduced false-premise questions to test critical thinking, while MMCV (Wang et al., 2025) extended verification into the multimodal domain. MUSER (Liao et al., 2023) and MSynFD (Xiao et al., 2024) introduced multi-step retrieval and syntax-aware mechanisms to handle complex narrative structures. Similarly, EX-FEVER (Ma et al., 2024a) emphasized explainability, requiring models to generate reasoning paths. While these works represent a significant leap forward in task complexity, they still rely on static data snapshots, leaving them vulnerable to the rapid obsolescence and contamination inherent in fixed datasets.

Despite these advancements, we face a "validity crisis" stemming from BDC and the lack of temporal realism (Xu et al., 2025a; Zhang et al., 2024a). Static benchmarks cannot guarantee that a model is reasoning on unseen data and do not capture the temporal constraints of real-world verification (Walter et al., 2020).

Recently, AdvFake (Chen et al., 2025a) tried to address BDC problem dynamically through real-time adversarial generation to stress-test RAG systems. While sharing our dynamic philosophy, LiveFact differs fundamentally: rather than focusing on adversarial deception, we isolate temporal reasoning and epistemic humility through strictly controlled time-sliced evidence (the "Fog of War").

The benchmark most closely aligned with Live-

Fact’s philosophy is LiveBench<sup>3</sup> (White et al., 2025), which uses a regularly updated design to mitigate contamination. However, LiveBench is not designed for fake news detection; instead, it evaluates general LLM capabilities such as coding, mathematics, and data analysis. Consequently, it lacks the specific evidential reasoning structures required to verify complex claims in the misinformation domain. In the realm of fake news detection, the most recent attempt is TripleFact (Xu and Yan, 2025). However, it remains a highly limited conceptual framework. Due to strict data copyright issues, the TripleFact dataset cannot be publicly released, and it lacks critical components such as continuous updates and temporal evidence slicing. Additionally, other recent benchmarks like FreeEval (Yu et al., 2024) and TreeEval (Li et al., 2025c) incorporate advanced evaluation methodologies, but they do not address the specific constraints of the fake news detection domain.

To the best of our knowledge, as shown in Table 1, LiveFact is the first fake news detection benchmark to simultaneously address dynamic updates, contamination control, and temporal reasoning, filling the critical gaps left by existing works.

### 3 The LiveFact Benchmark

#### 3.1 Motivation

LiveFact is designed around three core principles that address the limitations identified in Section 2. First, to combat BDC, we implement *continuous monthly updates* using the latest news events, en-

<sup>3</sup><https://livebench.ai/>

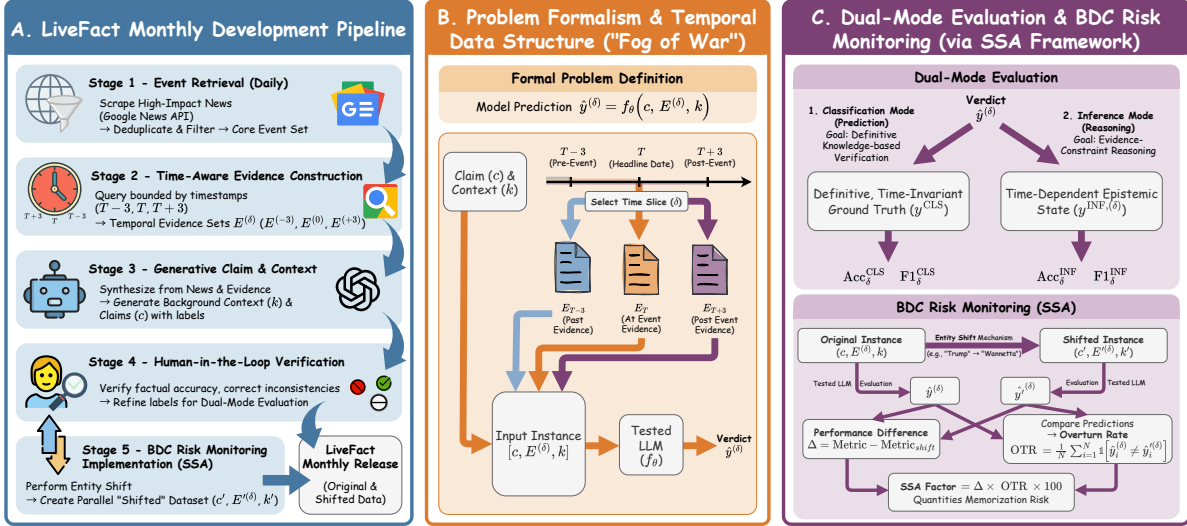


Figure 2: The overall framework of the LiveFact Benchmark. (A) The Monthly Development Pipeline illustrates the continuous process of acquiring real-time events, generating claims and context via LLMs, and performing human verification. (B) The Problem Formalism specifies the task as conditional reasoning under temporal constraints, utilizing time-sliced evidence sets (e.g., Pre-Event  $E^{(-3)}$  vs. Post-Event  $E^{(+3)}$ ) to simulate the "fog of war." (C) The Evaluation framework details the Dual-Mode approach (separating prediction vs. reasoning capabilities) and the integration of the SSA Framework to quantify BDC risk via entity shift mechanism.

During evaluation data remain unseen during model pre-training. Second, to assess genuine reasoning rather than knowledge retrieval, we introduce a *temporal evidence structure* that simulates the "fog of war"—providing models with evidence sets of varying completeness ( $E^{(\delta)}$ ) to test their ability to recognize information gaps. Third, to disentangle reasoning from memorization, we employ a *dual-mode evaluation*: Classification Mode tests factual accuracy against ground truth, while Inference Mode tests whether models appropriately predict "Ambiguous" when evidence is insufficient. These principles are formalized in the following section.

### 3.2 Problem Definition

We formalize the LiveFact task as a conditional reasoning problem under temporal constraints. Let  $\mathcal{E} = \{e_1, e_2, \dots, e_M\}$  denote a set of  $M$  news events, where each event  $e_j$  is associated with a headline date  $T_j$ . For each event  $e_j$ , we construct a set of claims  $\mathcal{C}_j = \{c_{j,1}, c_{j,2}, \dots\}$ , where each claim  $c \in \mathcal{C}_j$  is a statement about event  $e_j$ . We denote the full dataset as  $\mathcal{D} = \{(c_i, e(c_i), k_i, y_i)\}_{i=1}^N$ , consisting of  $N$  claim instances, where  $e(c_i)$  denotes the event associated with claim  $c_i$ ,  $k_i$  is the static background context describing the fundamental entities involved (e.g., "Donald Trump is the current President of the United States..."), and  $y_i \in \mathcal{Y}$  is the ground-truth label with  $\mathcal{Y} = \{\text{REAL}, \text{FAKE}, \text{AMBIGUOUS}\}$ .

To simulate the "fog of war" inherent in real-time verification, we construct time-sliced evidence sets based on the headline date of each event. For an event with headline date  $T$ , we define  $E^{(\delta)}$  as the evidence set containing only information available up to day  $T + \delta$ , where  $\delta \in \{-3, 0, +3\}$  represents the temporal offset in days. For instance,  $E^{(-3)}$  contains evidence published at least three days before the event,  $E^{(0)}$  includes evidence up to the headline date, and  $E^{(+3)}$  incorporates evidence available three days after the event.

Given a LLMs  $f_\theta$  parameterized by  $\theta$ , the model predicts a verdict based on the claim, time-sliced evidence, and background context:

$$\hat{y}_i^{(\delta)} = f_\theta(c_i, E_i^{(\delta)}, k_i) \quad (1)$$

where  $E_i^{(\delta)}$  denotes the evidence set at temporal offset  $\delta$  for the event associated with claim  $c_i$ , and  $\hat{y}_i^{(\delta)} \in \mathcal{Y}$  is the predicted label.

We evaluate models under two distinct modes:

- **Classification Mode** ( $m = \text{CLS}$ ): The ground-truth label  $y_i^{\text{CLS}}$  is the definitive, time-invariant verification label representing the ultimate factual status of the claim.
- **Inference Mode** ( $m = \text{INF}$ ): The ground-truth label  $y_i^{\text{INF},(\delta)}$  is dynamic, determined by whether the partial evidence  $E_i^{(\delta)}$  is sufficient

to support a correct judgment at the given temporal offset.

Let  $m \in \{\text{CLS}, \text{INF}\}$  denote the evaluation mode, and let  $y_i^{m,(\delta)}$  denote the corresponding ground-truth label (where  $y_i^{\text{CLS},(\delta)} = y_i^{\text{CLS}}$  is time-invariant for Classification Mode). We measure performance using accuracy and macro-averaged F1 score:

$$\text{Acc}_\delta^m = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[\hat{y}_i^{(\delta)} = y_i^{m,(\delta)}] \quad (2)$$

$$\text{F1}_\delta^m = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \frac{2 \cdot P_y^{m,(\delta)} \cdot R_y^{m,(\delta)}}{P_y^{m,(\delta)} + R_y^{m,(\delta)}} \quad (3)$$

where  $\mathbb{1}[\cdot]$  is the indicator function that returns 1 if the condition is true and 0 otherwise, and  $P_y^{m,(\delta)}$  and  $R_y^{m,(\delta)}$  denote precision and recall for class  $y \in \mathcal{Y}$  under mode  $m$  at temporal offset  $\delta$ , respectively.

To monitor the risk of BDC, we integrate the SSA framework (Xu et al., 2025b). For each instance, we generate a counterfactual claim  $c'_i$  via an Entity Shift mechanism, along with corresponding shifted evidence  $E'_i^{(\delta)}$  and context  $k'_i$ . Key entities are replaced with semantically equivalent but fictional alternatives (e.g., "Trump"  $\rightarrow$  "Wannetta") to isolate reasoning from memorization. Let  $\hat{y}'_i^{(\delta)} = f_\theta(c'_i, E'_i^{(\delta)}, k'_i)$  denote the prediction on the shifted instance.

We define the *Overtake Rate* (OTR) as the proportion of instances where the model’s prediction changes solely due to entity replacement:

$$\text{OTR} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[\hat{y}_i^{(\delta)} \neq \hat{y}'_i^{(\delta)}] \quad (4)$$

Let  $\Delta$  denote the performance gap between the original and shifted evaluations:

$$\Delta = \text{Metric} - \text{Metric}_{\text{shift}} \quad (5)$$

where Metric refers to any evaluation metric (e.g., accuracy or F1 score). The final contamination indicator is:

$$\text{SSA Factor} = \Delta \times \text{OTR} \times 100 \quad (6)$$

A high SSA-Factor indicates that the model relies on spurious correlations or memorized entity-specific knowledge rather than robust evidence-based reasoning.

### 3.3 LiveFact Development Pipeline

As shown in Figure 2, the construction of LiveFact is a rigorous, multi-stage process designed to ensure data freshness, accuracy, and depth. The pipeline is automated for monthly updates while incorporating human oversight to guarantee quality. For this inaugural iteration, our data collection spanned November 2025.

#### 3.3.1 Stage 1: Event Retrieval

We initiate the pipeline by collecting the latest high-impact news events. Using the Google News API, we systematically scrape trending events from the "World" section daily at 00:00 GMT. This ensures a consistent global scope for our dataset. To maintain dataset quality and reduce redundancy, we apply a rigorous deduplication and filtering process, removing overlapping storylines and low-relevance items. For the November 2025 cycle, this process yielded a core Event Set of 737 distinct events.

#### 3.3.2 Stage 2: Time-Aware Evidence Construction

To support our temporal reasoning analysis ("Fog of War"), we construct evidence sets anchored to specific time offsets relative to the event’s headline date ( $T$ ). We specifically target three distinct temporal offsets:  $\delta = -3$  (three days prior),  $\delta = 0$  (the event headline date), and  $\delta = +3$  (three days post-event). Leveraging Google APIs, we query for information strictly bounded by these timestamps, resulting in three discrete evidence sets for each event:  $E^{(-3)}$ ,  $E^{(0)}$ , and  $E^{(+3)}$ . In total, we gathered 25,064 distinct pieces of evidence across the 737 events, providing a rich informational backdrop for temporal analysis. Further technical specifications are detailed in Appendix A.1.

#### 3.3.3 Stage 3: Claim and Context Generation

Once the raw data are aggregated, we employ OpenAI’s o4-mini<sup>4</sup>, one of the most advanced reasoning models available, to synthesize the testing components. We input the main news headline and all associated time-stamped evidence into the model to generate two key outputs: (1) a concise background Context for the entities involved, and (2) a set of Claims categorized into three labels: Real, Ambiguous, and Fake. For the current dataset, this stage produced a total of 4,392 test sample claims, distributed initially as: "Real" (1,468), "Fake" (1,451),

<sup>4</sup><https://openai.com/index/introducing-o3-and-o4-mini/>

and "Ambiguous" (1,473). This balanced generation ensures the model is tested against diverse truth values. Detailed generation prompts and parameters are provided in Appendix A.2.

### 3.3.4 Stage 4: Human-in-the-Loop Verification

To ensure the reliability of our ground truth, all generated content undergoes a strict human review process. Annotators verify the factual accuracy of the generated claims against the retrieved evidence and correct any inconsistencies.

Crucially, this stage involves refining the labels for our Inference Mode. In this mode, the ground truth is dynamically adjusted based on whether the provided evidence slice ( $E^{(-3)}$ ,  $E^{(0)}$ , or  $E^{(+3)}$ ) is sufficient for a human expert to make a correct judgment. For instance, a claim that is factually "Real" might be labeled "Ambiguous" at  $T - 3$  if the confirming evidence had not yet emerged. This adjustment results in significant shifts in label distribution across time slices, as shown in Table 2.

Evidence Set	Real	Fake	Ambiguous	Total Claims
$E^{(-3)}$	346	348	3,698	4,392
$E^{(0)}$	1,404	1,390	1,598	4,392
$E^{(+3)}$	1,460	1,443	1,489	4,392

Table 2: Dynamic Label Distribution for Inference Mode (November 2025)

As illustrated, the  $\delta = -3$  set is heavily skewed toward "Ambiguous," reflecting the high uncertainty of the "Fog of War" period. As time progresses to  $\delta = 0$  and  $\delta = +3$ , the distribution normalizes as more evidence becomes available. This evolving distribution is a key feature of LiveFact, designed to test temporal adaptability. The complete protocol for manual review is documented in Appendix A.2.

### 3.3.5 Stage 5: BDC Risk Monitoring

Finally, we integrate the contamination monitoring component by adopting the workflow from the SSA framework (Xu et al., 2025b). To avoid over-reliance on OpenAI models that could lead to potential preference leakage (Li et al., 2025a), we employ the Qwen3-235B-A22B model (Yang et al., 2025) at this stage, we perform Entity Shift on the finalized Event Set, Claim Set, and Evidence Set. This process replaces specific named entities with fictional or neutral counterparts while preserving the narrative structure. By evaluating models on

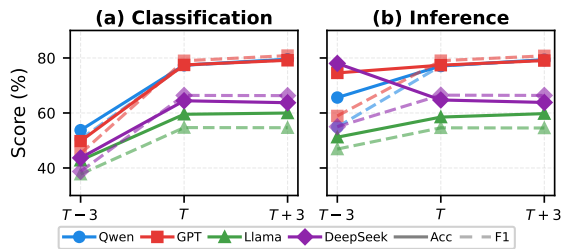


Figure 3: Temporal performance evolution for select model families across  $\delta \in \{-3, 0, +3\}$ . Panel (a) shows Classification Mode, where scores drop at  $\delta = -3$  due to the lack of definitive evidence. Panel (b) shows Inference Mode, where robust models recover accuracy by correctly predicting "Ambiguous," narrowing the performance gap.

this parallel "shifted" dataset, we calculate the performance difference and OTR to derive the SSA Factor, providing a quantitative measure of BDC risk for each release cycle. Further details regarding this stage are provided in Appendix C.1.

## 3.4 Evaluation Metrics

To rigorously evaluate model performance across different modes and time slices, we employ a comprehensive set of 12 distinct metrics.

We calculate Accuracy (Acc) and F1-Macro Score (F1) for both evaluation modes (Classification,  $m = \text{CLS}$ , and Inference,  $m = \text{INF}$ ) at each of the three temporal offsets ( $\delta \in \{-3, 0, +3\}$ ).

While Accuracy provides a general view of performance, it can be misleading when classes are imbalanced. As demonstrated in Stage 4 (Table 2), the manual adjustment for Inference Mode creates a severe class imbalance at  $\delta = -3$ , where "Ambiguous" labels account for nearly 85% of the data. A model that simply guesses "Ambiguous" for every query at  $\delta = -3$  would achieve a high accuracy score without performing any genuine reasoning. The F1-Macro score treats all classes equally, preventing the majority class from dominating the score. This allows us to rationally evaluate whether a model can correctly identify the rare "Real" or "Fake" instances that are verifiable at early stages, ensuring a robust assessment of reasoning capabilities under unbalanced conditions.

## 4 Experiments

### 4.1 Experiment Setup

For our main experimental evaluation, we selected 18 representative models to ensure broad cover-

Model	Eval. Mode	Classification						Inference						Avg.
		Acc <sub>-3</sub> <sup>CLS</sup>	Acc <sub>0</sub> <sup>CLS</sup>	Acc <sub>+3</sub> <sup>CLS</sup>	F1 <sub>-3</sub> <sup>CLS</sup>	F1 <sub>0</sub> <sup>CLS</sup>	F1 <sub>+3</sub> <sup>CLS</sup>	Acc <sub>-3</sub> <sup>INF</sup>	Acc <sub>0</sub> <sup>INF</sup>	Acc <sub>+3</sub> <sup>INF</sup>	F1 <sub>-3</sub> <sup>INF</sup>	F1 <sub>0</sub> <sup>INF</sup>	F1 <sub>+3</sub> <sup>INF</sup>	
Qwen3-235B-A22B-Instruct-2507		52.25	<u>79.76</u>	<b>82.08</b>	48.63	<b>79.84</b>	<u>82.02</u>	66.67	<b>79.53</b>	<b>81.99</b>	54.47	<u>79.64</u>	81.94	<b>72.40</b>
Qwen3-30B-A3B-Instruct-2507		<u>55.24</u>	75.05	77.00	<u>52.39</u>	74.87	76.61	64.55	74.64	76.78	55.37	74.66	76.40	69.46
Qwen3-32B		45.20	70.54	71.86	39.10	68.99	69.93	<b>82.81</b>	71.47	71.86	<u>60.94</u>	69.70	69.91	66.03
Qwen3-8B		50.84	68.88	69.83	49.12	67.79	68.44	60.29	68.67	69.65	53.76	67.82	68.30	63.62
Qwen3-4B-Instruct-2507		45.58	62.20	64.80	38.62	61.84	64.33	37.75	61.57	64.66	38.10	61.81	64.27	55.46
Llama-3.3-70B-Instruct		44.90	69.01	69.76	38.96	63.71	64.17	23.70	66.85	69.44	33.50	62.32	63.93	55.85
Llama-3.1-70B <sup>†</sup>		34.77	33.45	33.47	19.52	16.75	16.80	7.90	31.99	33.29	5.10	16.20	16.73	22.16
Llama-3.1-8B-Instruct		40.69	50.05	50.20	36.53	45.67	45.11	78.69	50.16	50.09	60.26	46.89	45.14	49.96
Llama-3.2-3B <sup>†</sup>		33.42	33.42	33.42	16.70	16.70	16.70	7.88	31.97	33.24	4.87	16.15	16.63	22.76
Llama-3.2-1B <sup>†</sup>		33.38	33.31	33.31	16.74	16.72	16.69	7.90	31.85	33.13	4.89	16.16	16.62	21.72
DeepSeek-V3.1		43.67	64.44	63.73	38.73	66.39	66.31	78.03	64.73	63.83	55.01	66.48	66.38	61.48
Kimi-K2-Thinking*		45.97	57.15	54.21	45.34	61.71	59.22	58.13	56.56	54.08	42.07	61.57	59.15	54.60
gpt-oss-120b*		<b>55.83</b>	<b>79.94</b>	<u>81.81</u>	<b>52.82</b>	<u>79.79</u>	81.49	62.23	<u>78.89</u>	<u>81.60</u>	50.83	78.97	81.31	<u>72.13</u>
gpt-oss-20b*		47.84	65.55	67.42	47.16	61.33	62.32	41.46	64.32	67.28	36.53	61.07	62.34	57.05
gpt-5.2-2025-12-11		47.63	76.34	77.32	42.88	78.20	79.56	<u>80.71</u>	76.25	77.30	<b>64.15</b>	78.30	79.57	71.52
gpt-5.1-2025-11-13		51.89	78.60	81.01	47.66	79.76	<b>82.17</b>	68.44	78.51	80.83	53.68	<b>79.72</b>	<b>81.99</b>	72.02
gpt-4o-2024-08-06		48.84	72.29	73.98	44.45	71.38	72.83	74.61	72.95	74.13	54.99	71.94	72.98	67.11
gpt-4o-mini-2024-07-18		51.94	71.02	71.22	49.06	69.54	69.43	73.86	71.17	71.22	55.92	69.67	69.45	66.12

Table 3: Performance comparison of 18 LLMs on LiveFact (November 2025). Models marked with <sup>†</sup> are base models without instruction tuning (detailed in Table 5). The best result in each column is **bolded**, and the second-best result is underlined. The last column, Avg., represents the average of all results tested for this model.  $E^{-3}$  CLS evaluates models ( $\text{Acc}_{-3}^{\text{CLS}}$ ,  $\text{F1}_{-3}^{\text{CLS}}$ ) under forced-choice conditions prior to event occurrence, highlighting prior bias rather than evidence-based verification. The asterisk \* denotes the non-standard evaluation setting required to accommodate verbose Chain-of-Thought (CoT) outputs.

age of current LLM capabilities. This selection includes 14 open-source models (spanning the Qwen3, Llama 3, DeepSeek, Kimi, and GPT-OSS families) and 4 closed-source proprietary models (from the OpenAI GPT series). This diverse cohort ranges in scale from 1 billion to 1 trillion parameters, covering everything from lightweight edge-deployable models to flagship commercial systems. The specific architectural details, parameter counts, and training methodologies for all 18 models are provided in Appendix B.1. All models were evaluated on the LiveFact November 2025 dataset. Detailed evaluation settings, including inference hyperparameters and prompt structures, are outlined in Appendix B.2.

## 4.2 General Results & Analysis

Table 3 presents the comprehensive performance of all 18 models across both Classification and Inference modes. The experimental results demonstrate a significant shift in the competitive landscape of LLMs for complex reasoning tasks. The Qwen3-235B-A22B-Instruct model not only leads the open-source sector but also outperforms proprietary flagship models, including gpt-5.1, achieving the highest average score of 72.40%. This dominance of Mixture-of-Experts (MoE) architectures—seen also in the strong performance of DeepSeek-V3.1 (61.48%) despite its massive pa-

rameter count—suggests that sparse, high-capacity models are particularly well-suited for the multifaceted nature of fake news detection, where knowledge retrieval must be dynamically routed. In contrast, traditional dense models like Llama-3.3-70B lag behind their MoE counterparts, highlighting potential limitations in scaling dense architectures for nuanced reasoning tasks.

It is important to clarify the interpretation of the Classification mode accuracy at the early  $E^{(-3)}$  slice. Because sufficient evidence to definitively verify a claim rarely exists before an event occurs, forcing a model to make a binary (Real/Fake) prediction at this stage acts as a "Hallucination Stress Test." A high CLS score at  $E^{(-3)}$  does *not* necessarily indicate superior reasoning. Instead, it often implies that the model relies heavily on its parametric memory, prior biases, or "clairvoyant guessing" based on the linguistic plausibility of the claim. Therefore, we include the  $E^{(-3)}$  CLS scores in the overall average to provide a holistic measure of a model’s forced-choice behavior across all constraints, but it should not be viewed as a standalone metric of fact-checking capability. The true measure of evidence-based reasoning is whether a model correctly transitions to the "Ambiguous" label in the Inference mode when faced with this same  $E^{(-3)}$  slice.

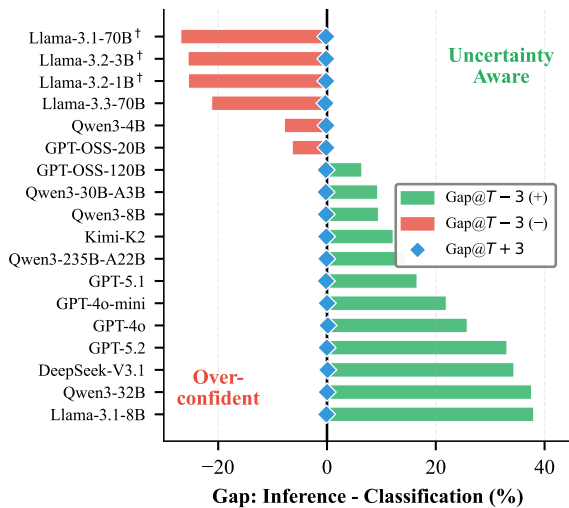


Figure 4: Analysis of the "Reasoning Gap" (Inference Score – Classification Score at  $\delta = -3$ ). Models with large positive gaps (green bars) effectively detect information voids ("Uncertainty Aware"). Models with negative or near-zero gaps (red bars) exhibit two failure modes: instruction-tuned models are genuinely "Overconfident," while base models ( $\dagger$ ) fail due to format non-compliance rather than reasoning deficits.

#### 4.2.1 Instruction Tuning and Reliability

The results highlight a critical performance distinction between standard instruction-tuned models and specialized "reasoning" models (\*). Base models ( $\dagger$ ) consistently fail to navigate the complex formatting and reasoning constraints of the benchmark due to structural non-compliance. A more nuanced finding, however, relates to models trained with extensive CoT reinforcement, such as Kimi-K2 and the GPT-OSS series. Initially, these models appeared to fail under standard evaluation settings (128 tokens), yielding near-zero accuracy due to truncated outputs. However, when re-evaluated with an extended limit of 1024 tokens, their performance rebounded dramatically, with gpt-oss-120b achieving a competitive 72.13% average. This suggests that for these architectures, "thinking" is not an optional feature but a structural necessity; they cannot be forced into immediate compliance without sacrificing their reasoning capabilities. We explore the inherent trade-offs between this "Thinking Mode" and standard instruction following in greater depth in Appendix D.

#### 4.2.2 Validating the LiveFact Paradigm: The "Reasoning Gap"

A core contribution of LiveFact is its ability to differentiate between simple hallucination and gen-

uine evidential reasoning. This differentiation is visualized in Figure 3, which tracks performance across the  $\delta = -3 \rightarrow \delta = 0 \rightarrow \delta = +3$  timeline. In Classification Mode, models suffer a severe performance drop at  $\delta = -3$  because they are forced to predict definitive labels like "Real" even when evidence is absent. However, in Inference Mode, where the ground truth is adjusted to "Ambiguous" to reflect the lack of evidence, we see a massive recovery in scores for capable models.

This phenomenon is further quantified in Figure 4, which plots the "Reasoning Gap" (Inference Accuracy minus Classification Accuracy at  $\delta = -3$ ). Models with a large positive gap (e.g., Llama-3.1-8B-Instruct with +38% and Qwen3-32B with +37%) can be characterized as *Uncertainty Aware*. They correctly identify that the evidence is insufficient and predict "Ambiguous," aligning with the Inference ground truth.

In contrast, models with negative or near-zero gaps fall into two distinct failure modes. *Instruction-tuned models* such as Llama-3.3-70B-Instruct and Qwen3-4B exhibit genuine *Overconfidence*: they produce valid classification outputs but continue to hallucinate "Real" or "Fake" verdicts regardless of the evidence void. *Base models* (marked with  $\dagger$  in Table 3), however, represent a fundamentally different failure mode. As discussed in Section 4.2.1, models like Llama-3.1-70B lack instruction tuning and fail primarily due to *structural non-compliance*—they cannot adhere to the required output format ([[LABEL]]), resulting in near-random parsed predictions rather than reasoned overconfidence. This distinction is critical: while overconfident models require improved calibration, base models require alignment training before their reasoning capabilities can be meaningfully assessed. A detailed confusion matrix analysis in Appendix D.3 further corroborates these findings, revealing that high-performing models exhibit clear "diagonalization" at  $\delta = 0$  and  $\delta = +3$ , while base models show pathological single-class collapse.

#### 4.2.3 Efficiency vs. Capability Analysis

We further analyze the relationship between model scale, cost, and performance using Figure 1. The figure plots the "Efficiency Frontier," revealing a non-linear relationship between resource consumption and utility. While massive models like Qwen3-235B-A22B and gpt-5.2 define the state-of-the-art (SOTA), they incur high operational costs (\$3.65

- \$9.27 per run). A compelling "efficiency sweet spot" emerges with mid-sized MoE models. As highlighted in the figure, Qwen3-30B-Instruct delivers 69.46% average performance—within 3 percentage points of the leader—at approximately 14x cheaper cost than gpt-5.2. This suggests that for high-volume, real-time fake news detection applications, well-tuned mid-sized models offer a superior ROI compared to giant foundation models.

## 5 Conclusion

In this paper, we introduced LiveFact, a dynamic, time-aware benchmark designed to address the critical limitations of static fake news detection evaluations in the era of LLMs. By shifting the paradigm from simple knowledge retrieval to temporal reasoning under the "Fog of War," LiveFact provides a rigorous testbed for assessing whether models are genuinely reasoning or merely recalling pre-trained data. Our extensive experiments with 18 diverse LLMs reveal that while SOTA models exhibit strong reasoning capabilities, significant gaps remain in "epistemic humility" and instruction adherence, particularly among base models. Furthermore, we established that mid-sized MoE architectures offer a compelling balance of efficiency and performance. Finally, while the full utility of SSA will be realized through LiveFact's monthly updates to monitor long-term BDC risk, we have validated its mechanism through a simulated BDC injection experiment detailed in Appendix C. To facilitate community engagement and track ongoing progress in LLM verification capabilities, we maintain an interactive, regularly updated leaderboard<sup>5</sup>. We hope LiveFact serves as a sustainable gold standard, pushing the field toward more robust, transparent, and temporally grounded AI verification systems.

## Limitations

While LiveFact represents a significant advancement in dynamic, time-aware evaluation, we identify three primary limitations in its current iteration. First, the current benchmark is predominantly constructed using English-language sources from global news outlets. Consequently, it may not fully capture the cultural nuances, localized disinformation tactics, or linguistic complexities present in non-English regions, where misinformation often proliferates rapidly (Shibu et al., 2025).

<sup>5</sup><https://livefact.bebxy.com/>

Future work will aim to expand LiveFact into a multilingual framework, incorporating diverse regional sources to test cross-cultural reasoning capabilities. Second, our current framework focuses exclusively on textual claims and evidence. However, modern misinformation increasingly relies on multimodal content, including manipulated images, deepfakes, and decontextualized videos (Jing et al., 2023; Khattar et al., 2019; Hu et al., 2025b). While text-based reasoning is foundational, it represents only one facet of the problem. We plan to extend future iterations of LiveFact to include multimodal evidence retrieval and verification tasks, aligning with the broader trajectory of LVLM (Large Vision-Language Model) development. Third, although our human-in-the-loop pipeline ensures high-quality ground truth (review synthetic data and ensure semantic consistency of entity shift data), it introduces a bottleneck regarding scalability. The requirement for expert review to distinguish between "Ambiguous" and "False" labels in Inference Mode limits the volume of claims we can release in each monthly cycle compared to fully synthetic datasets. To address this, we plan to investigate semi-automated verification protocols where reliable "Judge" models—calibrated against our human-verified corpus—can assist in the review process, thereby increasing dataset throughput without compromising rigor.

## Ethical Considerations

We strictly adhere to ethical guidelines regarding data privacy, copyright compliance, and model usage throughout the development of LiveFact. The application of AI assistants is limited solely to enhancing paper writing. All LLMs utilized in this study were accessed in strict accordance with their respective usage policies. For open-source models (e.g., Llama 3, Qwen3, DeepSeek), we utilized official checkpoints released under their specific community licenses. For proprietary models (e.g., OpenAI GPT series), access was obtained through authorized API endpoints, complying with all terms of service regarding automated evaluation and benchmarking. Our evidence retrieval pipeline relies on the Google API. To rigorously respect copyright regulations and intellectual property rights, we do not scrape, store, or redistribute the full text of news articles. As detailed in Appendix A.1, our dataset retains only minimal metadata—specifically webpage titles, source names,

and publication timestamps—which falls within standard fair use for research purposes, also the data does not contain any privacy-sensitive data beyond public information. This metadata is sufficient for models to perform retrieval-augmented reasoning without infringing on publisher content. All data and artifacts produced in LiveFact are intended solely for academic research to enhance the safety and reliability of AI systems. We acknowledge the dual-use nature of fake news research (Sun et al., 2024; Hu et al., 2025a); however, our focus remains strictly on detection and verification. In all future long-term updates of the LiveFact benchmark, we commit to maintaining this rigorous standard of compliance, ensuring that all data sourcing and model interactions continue to adhere to evolving legal frameworks and ethical agreements.

## Acknowledgments

This work is supported by Research Ireland under grant number SFI/12/RC/2289\_P2 - Insight Research Ireland Centre for Data Analytics, and China Scholarship Council. We also acknowledge the support from Thinking Machines Lab and OpenAI Inc. for this work.

## References

- Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. [Where is your evidence: Improving fact-checking by justification modeling](#). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.
- Anthropic. 2024. [Introducing the next generation of claude](#).
- Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. [Generating fact checking explanations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.
- Han Cao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2025. [Enhancing multi-hop fact verification with structured knowledge-augmented large language models](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23514–23522.
- Jianhong Chen, Wenyi Zhang, Hongcai Ma, and Shan Yang. 2023. [Rumor detection in social media based on multi-hop graphs and differential time series](#). *Mathematics*, 11(16).
- Sanxing Chen, Yukun Huang, and Bhuwan Dhingra. 2025a. [Real-time factuality assessment from adversarial feedback](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1610–1630, Vienna, Austria. Association for Computational Linguistics.
- Simin Chen, Yiming Chen, Zexin Li, Yifan Jiang, Zhongwei Wan, Yixin He, Dezhi Ran, Tianle Gu, Haizhou Li, Tao Xie, and Baishakhi Ray. 2025b. [Recent advances in large language model benchmarks against data contamination: From static to dynamic evaluation](#). *Preprint*, arXiv:2502.17521.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Chunyan 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 and virtual meeting. Association for Computational Linguistics.
- Shuzhi Gong, Richard O. Sinnott, Jianzhong Qi, and Cecile Paris. 2023. [Fake news detection through graph-based neural networks: A survey](#). *Preprint*, arXiv:2307.12639.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 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.
- Alexander Havrilla, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, and Roberta Raileanu. 2024. [GLore: When, where, and how to improve LLM reasoning via global and local refinements](#). In *Forty-first International Conference on Machine Learning*.
- Beizhe Hu, Qiang Sheng, Juan Cao, Yang Li, and Danding Wang. 2025a. [Llm-generated fake news induces truth decay in news ecosystem: A case study on neural news recommendation](#). In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’25*, page 435–445, New York, NY, USA. Association for Computing Machinery.
- Lingtong Hu, Zituo Wang, Jiayi Zhu, Yifan Hu, and Xianbing Wang. 2025b. [Mage-fend: Multimodal](#)

- adaptive fusion with guidance from llm expertise for fake news detection on short video platforms. *Knowledge-Based Systems*, 329:114298.
- Zhongqiang Huang, Dongli Lu, and Ying Sha. 2025. Multi-hop attention diffusion graph neural networks for multimodal fake news detection. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.
- Jing Jing, Hongchen Wu, Jie Sun, Xiaochang Fang, and Huaxiang Zhang. 2023. Multimodal fake news detection via progressive fusion networks. *Information Processing & Management*, 60(1):103120.
- Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference, WWW '19*, page 2915–2921, New York, NY, USA. Association for Computing Machinery.
- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.
- Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and huan liu. 2025a. Preference leakage: A contamination problem in LLM-as-a-judge. In *Data in Generative Models - The Bad, the Ugly, and the Greats*.
- Sihang Li, Jin Huang, Jiayi Zhuang, Yaorui Shi, Xiaochen Cai, Mingjun Xu, Xiang Wang, Linfeng Zhang, Guolin Ke, and Hengxing Cai. 2025b. Scilitlm: How to adapt llms for scientific literature understanding. In *The Thirteenth International Conference on Learning Representations*.
- Xiang Li, Yunshi Lan, and Chao Yang. 2025c. Treeeval: Benchmark-free evaluation of large language models through tree planning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24485–24493.
- Yupeng Li, Haorui He, Jin Bai, and Dacheng Wen. 2024. Mcfend: A multi-source benchmark dataset for chinese fake news detection. In *Proceedings of the ACM Web Conference 2024, WWW '24*, page 4018–4027, New York, NY, USA. Association for Computing Machinery.
- 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 '23*, page 4461–4472, New York, NY, USA. Association for Computing Machinery.
- Xuannan Liu, Zekun Li, Pei Pei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng, and Zhaofeng He. 2025. MMFakebench: A mixed-source multimodal misinformation detection benchmark for LVLMS. In *The Thirteenth International Conference on Learning Representations*.
- Yahui Liu, Wanlong Bing, Shuai Ren, and Hongliang Ma. 2024. Bc-fnd: An approach based on hierarchical bilinear fusion and multimodal consistency for fake news detection. *IEEE Access*, 12:62738–62749.
- 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.
- Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang Wang, Qiang Liu, Shu Wu, and Liang Wang. 2024a. EX-FEVER: A dataset for multi-hop explainable fact verification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9340–9353, Bangkok, Thailand. Association for Computational Linguistics.
- Xiaoxiao Ma, Yuchen Zhang, Kaize Ding, Jian Yang, Jia Wu, and Hao Fan. 2024b. On fake news detection with LLM enhanced semantics mining. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 508–521, Miami, Florida, USA. Association for Computational Linguistics.
- Timothy R. McIntosh, Teo Susnjak, Tong Liu, Paul Watters, and Malka N. Halgamuge. 2024. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *Preprint*, arXiv:2402.09880.
- Marion Meyers, Gerhard Weiss, and Gerasimos Spanakis. 2020. Fake news detection on twitter using propagation structures. In *Disinformation in Open Online Media*, pages 138–158, Cham. Springer International Publishing.
- Rahul Mishra. 2020. Fake news detection using higher-order user to user mutual-attention progression in propagation paths. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- Jeppe Nørregaard, Benjamin D. Horne, and Sibel Adalı. 2019. Nela-gt-2018: A large multi-labelled news

- dataset for the study of misinformation in news articles. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):630–638.
- OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. *gpt-oss-120b & gpt-oss-20b model card*. *Preprint*, arXiv:2508.10925.
- OpenAI. 2024. *Gpt-4 technical report*. *Preprint*, arXiv:2303.08774.
- Wojciech Ostrowski, Arnav Arora, Pepa Atanasova, and Isabelle Augenstein. 2021. *Multi-hop fact checking of political claims*. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3892–3898. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- 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.
- 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.
- Fariba Sadeghi, Amir Jalaly Bidgoly, and Hossein Amirkhani. 2022. *Fake news detection on social media using a natural language inference approach*. *Multimedia Tools and Applications*, 81(23):33801–33821.
- Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. *NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark*. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.
- Mohammadamin Shafiei, Hamidreza Saffari, and Nafise Sadat Moosavi. 2025. *MultiHoax: A dataset of multi-hop false-premise questions*. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10169–10187, Vienna, Austria. Association for Computational Linguistics.
- Hrithik Majumdar Shibu, Shrestha Datta, Md. Sumon Miah, Nasrullah Sami, Mahruba Sharmin Chowdhury, and Md Saiful Islam. 2025. *From scarcity to capability: Empowering fake news detection in low-resource languages with LLMs*. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 100–107, Abu Dhabi. Association for Computational Linguistics.
- 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. PMID: 32491943.
- Jiasheng Si, Yingjie Zhu, and Deyu Zhou. 2023. *Exploring faithful rationale for multi-hop fact verification via salience-aware graph learning*. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13573–13581.
- Yanshen Sun, Jianfeng He, Limeng Cui, Shuo Lei, and Chang-Tien Lu. 2024. *Exploring the deceptive power of llm-generated fake news: A study of real-world detection challenges*. *Preprint*, arXiv:2403.18249.
- 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*. In *Forty-second International Conference on Machine Learning*.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. *Stanford alpaca: An instruction-following llama model*. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca).
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, , and et al. 2024. *Gemini: A family of highly capable multimodal models*. *Preprint*, arXiv:2312.11805.
- Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, and 150 others. 2025. *Kimi k2: Open agentic intelligence*. *Preprint*, arXiv:2507.20534.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. *FEVER: a large-scale dataset for fact extraction and VERification*. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- Zhao Tong, Yimeng Gu, Huidong Liu, Qiang Liu, Shu Wu, Haichao Shi, and Xiao-Yu Zhang. 2025. *Generate first, then sample: Enhancing fake news detection with LLM-augmented reinforced sampling*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24276–24290, Vienna, Austria. Association for Computational Linguistics.

- Nathan Walter, Jonathan Cohen, R. Lance Holbert, and Yasmin Morag. 2020. [Fact-checking: A meta-analysis of what works and for whom](#). *Political Communication*, 37(3):350–375.
- Haoran Wang, Aman Rangapur, Xiong Xiao Xu, Yueqing Liang, Haroon Gharwi, Carl Yang, and Kai Shu. 2025. [Piecing it all together: Verifying multi-hop multimodal claims](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7453–7469, Abu Dhabi, UAE. 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.
- Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddhartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2025. [Livebench: A challenging, contamination-limited LLM benchmark](#). In *The Thirteenth International Conference on Learning Representations*.
- Liang Xiao, Chongyang Shi, Shufeng Hao, and Zeyu Wei. 2025. [Fact-augmented reasoning model for fake news detection](#). In *Neural Information Processing*, pages 18–32, Singapore. Springer Nature Singapore.
- Liang Xiao, Qi Zhang, Chongyang Shi, Shoujin Wang, Usman Naseem, and Liang Hu. 2024. [Msynfd: Multi-hop syntax aware fake news detection](#). In *Proceedings of the ACM Web Conference 2024, WWW ’24*, page 4128–4137, New York, NY, USA. Association for Computing Machinery.
- 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 and M-Tahar Kechadi. 2025. [Analysis of semantic benchmark data contamination attack for llm-driven fake news detection](#). In *2025 IEEE International Conference on Big Data (BigData)*, pages 3656–3664.
- Cheng Xu and Nan Yan. 2025. [TripleFact: Defending data contamination in the evaluation of LLM-driven fake news detection](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8808–8823, Vienna, Austria. Association for Computational Linguistics.
- Cheng Xu, Nan Yan, Shuhao Guan, Changhong Jin, Yuke Mei, Yibing Guo, and Tahar Kechadi. 2025a. [DCR: Quantifying data contamination in LLMs evaluation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23013–23031, Suzhou, China. Association for Computational Linguistics.
- Cheng Xu, Nan Yan, Shuhao Guan, Yuke Mei, and Tahar Kechadi. 2025b. [SSA: Semantic contamination of LLM-driven fake news detection](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14748–14762, Suzhou, China. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Chen Yang, Xinyi Zhou, and Reza Zafarani. 2021. [Checked: Chinese covid-19 fake news dataset](#). *Social Network Analysis and Mining*, 11(1):58.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.
- 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](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 1–13, Miami, Florida, USA. Association for Computational Linguistics.
- Chaowei Zhang, Zongling Feng, Zewei Zhang, Jipeng Qiang, Guandong Xu, and Yun Li. 2025. [Is llms hallucination usable? llm-based negative reasoning for fake news detection](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(1):1031–1039.
- Congzhi Zhang, Linhai Zhang, and Deyu Zhou. 2024a. [Causal walk: Debiasing multi-hop fact verification with front-door adjustment](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19533–19541.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. 2024b. [LLM as a mastermind: A survey of strategic reasoning with large language models](#). In *First Conference on Language Modeling*.

Congyuan Zhao, Lingwei Wei, Ziming Qin, Wei Zhou, Yunya Song, and Songlin Hu. 2025. [Mppfnd: A dataset and analysis of detecting fake news with multi-platform propagation](#). *Preprint*, arXiv:2505.15834.

Liwen Zheng, Chaozhuo Li, Litian Zhang, Haoran Jia, Senzhang Wang, Zheng Liu, and Xi Zhang. 2025. [Mrr-fv: Unlocking complex fact verification with multi-hop retrieval and reasoning](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):26066–26074.

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](#). *Preprint*, arXiv:2311.01964.

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

Ziyi Zhou, Xiaoming Zhang, Litian Zhang, Jiacheng Liu, Senzhang Wang, Zheng Liu, Xi Zhang, Chaozhuo Li, and Philip S. Yu. 2024. [Fine-fake: A knowledge-enriched dataset for fine-grained multi-domain fake news detection](#). *Preprint*, arXiv:2404.01336.

Junyou Zhu, Chao Gao, Ze Yin, Xianghua Li, and Juergen Kurths. 2024. [Propagation structure-aware graph transformer for robust and interpretable fake news detection](#). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 4652–4663, New York, NY, USA. Association for Computing Machinery.

## A LiveFact Development Details

### A.1 Temporal Evidence Sets Construction

For each event gathered in the initial collection stage, we performed a comprehensive evidence retrieval process using the Google API. Our goal was to build a robust informational environment for each event across different temporal slices. We sent targeted requests to the API and aggregated metadata from all search results. A typical JSON response structure for an evidence item is as follows:

```

1 {
2   "article_id": "1f5faebba75d590d",
3   "title": "COP30: Oil Rich Nations
         Block Deal To Phase Out Fossil
         Fuels",
4   "source": "Al-Fanar Media",

```

```

5   "published_datetime": "2025-11-27
6   T13:19:46Z"

```

The critical step in our pipeline involves the temporal segmentation of this data. By comparing the `published_datetime` of each retrieved page against the specific headline date ( $T$ ) of the event, we categorized the data into three distinct evidence buckets:

1.  $E^{(-3)}$ : Contains all evidence published up to three days before the headline date.
2.  $E^{(0)}$ : Contains all evidence published up to and including the headline date.
3.  $E^{(+3)}$ : Contains all evidence published up to three days after the headline date.

The choice of the 3-day window is informed by our empirical analysis of information velocity in digital news cycles. We observed that evidence density typically peaks within 48 to 72 hours relative to the headline date. Our pilot studies showed that extending the window to  $\pm 7$  or  $\pm 15$  days yielded diminishing returns: the periods of  $T - 15$  to  $T - 7$  often contained negligible relevant reporting, while the  $T + 7$  to  $T + 15$  intervals were characterized by redundant information or the 'cooling down' of the event. Conversely, narrower windows (e.g.,  $\pm 1$  day) risked missing crucial initial reports or early developments. Thus,  $T \pm 3$  days serves as an optimal balance to capture the 'Fog of War' without diluting the dataset with settled history. Furthermore, while minute-level granularity was considered, the prevalence of day-level precision in web metadata and the technical risk of timestamp leakage during crawling made daily-level increments more reliable and sustainable for long-term benchmark maintenance.

The rigorous 3-day segmentation resulted in the collection of 25,064 individual pieces of evidence across the 737 events, yielding a dense average of approximately 34 evidence items per event. This volume ensures that the models have sufficient context to perform complex reasoning rather than simple fact retrieval.

### A.2 Claim and Context Generation and Human Verification

To generate high-quality claims and context, we utilized two distinct LLMs to assess both performance and cost-efficiency: o4-mini and gpt-4o-mini from OpenAI. We supplied each model with the event

gpt-4o-mini	o4-mini
<p><b>(Context)</b> The Chinese astronauts are members of the China National Space Administration (CNSA), which is responsible for the country’s space program. Shenzhou-20 is a crewed spacecraft designed to transport astronauts to and from space, while Shenzhou-22 is an upcoming mission intended to provide a lifeboat for the stranded crew.</p>	<p><b>(Context)</b> Chinese astronauts are career space travelers trained and selected by the China National Space Administration (CNSA). The CNSA is China’s national space agency that operates the Shenzhou crewed spacecraft program, including the Shenzhou-20 and Shenzhou-22 missions to the Tiangong space station.</p>
<p><b>(Real)</b> Three Chinese astronauts are currently stranded in space after the successful return of their colleagues aboard the Shenzhou-20, with a replacement mission, <i>Shenzhou-22</i>, scheduled for launch on November 25 to provide a lifeboat for the stranded crew.</p>	<p><b>(Real)</b> After <i>space debris damaged the Shenzhou-20</i> return capsule, three Chinese astronauts remain temporarily stranded aboard China’s space station while Beijing prepares to launch the <i>uncrewed Shenzhou-22 lifeboat spacecraft</i> on November 25 to bring them home."</p>
<p><b>(Fake)</b> In a surprising turn of events, Chinese authorities have announced that <i>five additional astronauts have been successfully rescued</i> from the Shenzhou-20 mission, leaving <i>only two astronauts stranded</i> in space due to a malfunction in the rescue capsule, which is expected to be <i>operational again</i> by December 10.</p>	<p><b>(Fake)</b> A leaked internal bulletin from the China Manned Space Agency, obtained by Beijing News, reveals the three astronauts stranded in orbit have already <i>consumed over 65% of their emergency oxygen</i> and <i>suffered two failures in their primary water-recycling unit</i>, placing them at serious risk of critical life-support shortages before the delayed rescue craft arrives.</p>
<p><b>(Ambiguous)</b> Following the successful rescue of their colleagues, three more Chinese astronauts are now stranded in space, as <i>internal documents suggest</i> that the Chinese space agency is <i>prioritizing a secret mission</i> to test advanced life support systems over their immediate return, raising concerns among <i>unnamed experts</i> about the long-term implications for astronaut safety.</p>	<p><b>(Ambiguous)</b> China postponed the Shenzhou-22 docking by 45 minutes—officially to fine-tune orbital alignment—but the tweak was <i>actually aimed at slotting the lifeboat rendezvous with prime-time state TV bulletins</i>, underscoring Beijing’s <i>crafted image of flawless crisis management</i> even as three astronauts remained stranded.</p>

Table 4: Comparison of claim generation example between the gpt-4o-mini and o4-mini (Event ID: EV20251119-33). *Italicized* text highlights the specific factual anchors (Real), fabricated details (Fake), and unverifiable motives or sources (Ambiguous) that justify the assigned label.

headline and the full set of associated evidence, using specialized prompts to instruct them to generate: 1. **Context:** A neutral summary of the key entities involved; 2. **Claims:** A set of statements categorized as Real, Fake, or Ambiguous. The prompts used for these generations process were provide in Figure 5, 6, 7, 8

Following generation, we implemented a rigorous Human-in-the-Loop verification protocol. Every generated context and claim underwent a minimum of three rounds of independent review. The review team consisted exclusively of the paper’s authors, who are PhD-level computer science researchers specializing in fake news detection and natural language processing from English-speaking countries. The review criteria focused on factual consistency, linguistic naturalness, and label accuracy relative to the evidence. Any data point that failed to achieve consensus after three review rounds was discarded and regenerated until it met the quality standards. We observed a high pass rate (exceeding 90%), which can be attributed to our use of strict few-shot prompting templates that constrain the models to the factual structures of the news source. Given that generating short claims

from structured reporting is a relatively straightforward task for SOTA models like GPT-4o and GPT-4o-mini, the output quality was consistently high. The small fraction of rejected samples primarily fell under the "Ambiguous" category. For instance, a model-generated claim introducing an unverifiable micro-event (e.g., "António Guterres paused for exactly 12 seconds during the final COP30 plenary") caused disagreement regarding its phrasing bias, and was thus discarded to maintain strict benchmark objectivity.

We compared the two models to validate the generalizability and accessibility of the LiveFact framework:

- **gpt-4o-mini:** Selected to represent a highly versatile and cost-effective solution. The total cost for the generation process using this model was approximately 2.34 USD.
- **o4-mini:** Selected to represent a state-of-the-art inference-optimized model. The generation process for this model incurred a significantly higher cost of 55.93 USD.

Our analysis revealed that both models performed well on the task (exemplified in Table 4),

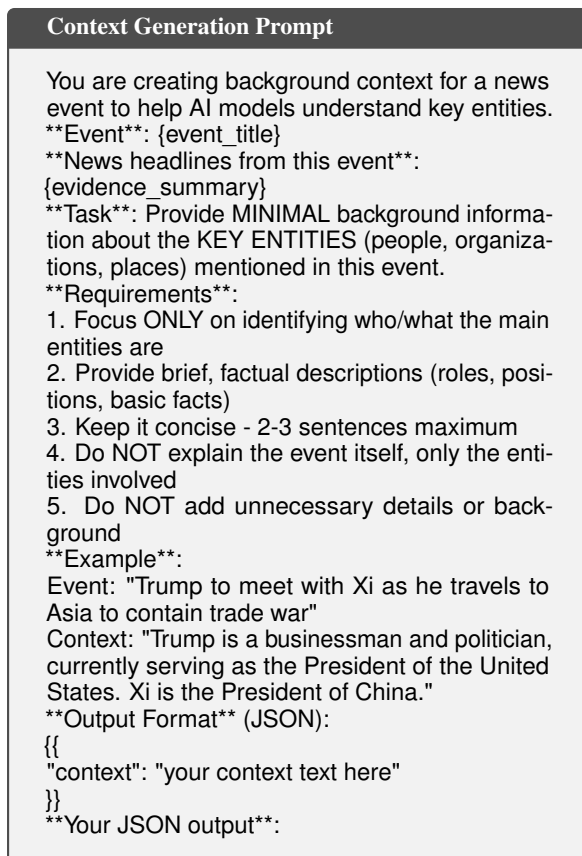


Figure 5: Prompt for Context Generation

producing high-quality initial outputs that required minimal manual correction. This finding validates the universality of the LiveFact construction pipeline, suggesting that future updates can be sustainably generated even with lower-cost models. However, to ensure the highest possible baseline for this inaugural release, we utilized the outputs from o4-mini for all subsequent experimental steps in this paper.

## B Experiment Details

### B.1 Model Details

To comprehensively evaluate the state of LLMs in 2025, we selected a representative cohort of 18 models spanning different architectures (Dense vs. MoE), training strategies (Base, Instruct, Hybrid), and parameter scales (1B to 1T). This set includes 14 open-source models and 4 closed-source proprietary models, categorized by family:

- **Qwen Family**: Represents the leading edge of open-source MoE performance, ranging from the massive Qwen3-235B-A22B to the compact, edge-ready Qwen3-4B (Yang et al., 2025).

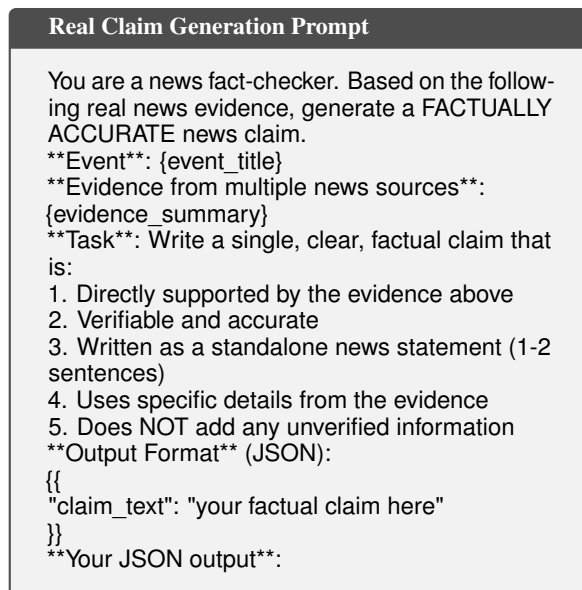


Figure 6: Prompt for Real Claim Generation

- **Llama Family**: Includes both the standard Dense architecture (Llama 3.1 70B/8B) and recent lightweight variants (Llama 3.2 3B/1B), serving as the baseline for open-weights performance (Grattafiori et al., 2024).
- **DeepSeek & Kimi**: Represent specialized massive-scale MoE models, with Kimi-K2 exceeding 1 trillion parameters (DeepSeek-AI et al., 2025; Team et al., 2025).
- **GPT-OSS Series**: This family represents high-performance open-weights implementations of GPT-style architectures, developed by decentralized research collectives. Models like gpt-oss-120b and gpt-oss-20b leverage advanced MoE structures to replicate reasoning capabilities similar to proprietary counterparts, offering a transparent benchmark for community-driven progress (OpenAI et al., 2025).
- **GPT Series (Proprietary)**: Serves as the commercial gold standard, including the latest gpt-5.2 and gpt-5.1 models alongside the efficient gpt-4o series (OpenAI, 2024).

Among the 18 open-source models available in our registry (Table 5), 14 were selected for the main experimental comparison, with the remaining 4 (primarily base model variants) used in supplementary ablation studies (see Appendix D). This selection ensures our results cover the full spectrum of modern AI capabilities, from cost-effective

```

Fake Claim Generation Prompt

You are creating synthetic fake news examples
for a research benchmark. Based on the event
below, generate a FABRICATED news claim.
**Event**: {event_title}
**Real evidence** (for context only - your claim
should CONTRADICT or FABRICATE beyond
this):
{evidence_summary}
**Task**: Write a fake news claim that is:
1. Related to the event but NOT supported by
evidence
2. Contains fabricated details (false numbers,
fake quotes, wrong people, etc.)
3. Sounds plausible but is factually incorrect
4. Written as a standalone news statement (1-2
sentences)
5. Could mislead someone unfamiliar with the
real story
**Important**: Make it realistic fake news, not
obviously absurd.
**Output Format** (JSON):
{{
"claim_text": "your fabricated claim here"
}}
**Your JSON output**:

```

Figure 7: Prompt for Fake Claim Generation

general-purpose models to the absolute peak of computational power.

## B.2 Evaluation Settings

To ensure a fair and reproducible evaluation, all models were subjected to a standardized inference protocol. We set the decoding parameters to `TEMPERATURE = 0.0` and `TOP_P = 1.0` to minimize randomness and ensure deterministic outputs. The maximum generation length was capped at `MAX_NEW_TOKENS = 128` (except for `Kimi-K2-Thinking` and `GPT-OSS`, they use `MAX_NEW_TOKENS = 1024`) to encourage concise verdicts. Note that for certain recent OpenAI models where these precise parameters are abstracted, we utilized the default "deterministic" settings provided by the API.

Models were prompted with a standardized template (Figure 9) instructing them to output their verdict in a strict parsable format: `[[LABEL]]` (e.g., `[[Real]]`, `[[Fake]]`, or `[[Ambiguous]]`).

All 14 open-source models were deployed and accessed via the Tinker<sup>6</sup> high-performance inference server, utilizing vLLM for optimized throughput. The 4 proprietary OpenAI models were ac-

<sup>6</sup><https://thinkingmachines.ai/tinker/>

cessed via standard OpenAI API<sup>7</sup> endpoints.

## B.3 Resource Consumption

Table 6 details the resource footprint for each model during the evaluation of the full LiveFact benchmark (across all three time slices). The data reveals striking disparities in the cost-to-performance ratio. High-end proprietary models like `gpt-5.2` incur the highest operational costs (\$9.27 per complete run), primarily due to their pricing per token. In contrast, the mid-sized open-source models demonstrate remarkable efficiency. For instance, `Qwen3-30B-A3B-Instruct` completes the benchmark for just \$0.64, which is approximately 14x cheaper than `gpt-5.2` while retaining competitive accuracy. This efficiency is largely driven by its MoE architecture, which activates only 3B parameters per token, drastically reducing computational overhead compared to dense counterparts like `Llama-3.1-70B` (\$5.22). This analysis suggests that for large-scale deployment, MoE architectures offer the optimal balance of reasoning capability and economic sustainability.

## C SSA Framework Analysis

### C.1 SSA Framework and Entity Shift Implementation Details

While the primary utility of the SSA framework lies in the long-term monitoring of historical benchmarks, we conducted a rigorous simulation experiment to validate its immediate efficacy in detecting label-level contamination. To ensure a robust experimental setup free from dependency on proprietary API models, we utilized the high-performing open-source model `Qwen3-235B-A22B-Instruct-2507` to execute the entity shift operations, prompt used in Figure 10. The quality of this synthetic data generation was strictly audited (consistent with Appendix A.2), and revealed that 98.5% of the generated entity-shifted samples were semantically consistent and required no further adjustment.

To simulate contamination, we constructed a poisoned training dataset derived from the LiveFact  $T + 3$  data split (Classification Mode). We mixed this specific benchmark data (formatted as instruction-response pairs) with a randomly selected subset of the Alpaca instruction fine-tuning dataset (Taori et al., 2023) (contaminants account for 30% of the total). This creates a scenario where the model has "seen" the exact answers to the test

<sup>7</sup><https://openai.com/api/>

### Ambiguous Claim Generation Prompt

You are creating synthetic data for a fact-checking benchmark. Your goal is to generate a "Hard-to-Verify" claim based on the provided event.

**\*\*Event\*\***: {event\_title} **\*\*Evidence\*\***: {evidence\_summary}

**\*\*Task\*\***: Generate a single-sentence news claim that is empirically hard to verify. The claim must anchor itself in verifiable facts but introduce ambiguity through interpretation, causality, or specific unrecorded details, sounds plausible but is factually hard to confirm. However, content that contradicts facts and evidence must not appear, as this is not about creating fake news.

**\*\*Constraint on Sources\*\***:

- **\*\*Avoid Over-reliance on Anonymous Sources\*\***: Do NOT start every claim with "Sources say" or "Insiders revealed." Only use attribution logic if it provides a specific context that cannot be expressed otherwise. Also, avoid using informal expressions like "quietly" or "privately," as this would make it too different from real news.  
- **\*\*Variety is Key\*\***: Prioritize the techniques below over generic hearsay.

**\*\*Required Techniques (Prioritize these over generic "insider" claims)\*\***:

1. **\*\*Imputed Intent\*\***: State a public figure's private, internal motivation for a verifiable action as an objective fact (e.g., "He signed the bill \*specifically to appease his donors\*").
2. **\*\*False Causality\*\***: Connect two unrelated real facts as direct cause-and-effect without proof (e.g., "The stock drop \*was a calculated response to\* the leaked memo").
3. **\*\*Subjective Framing\*\***: Use unverifiable, qualitative descriptors to characterize a verifiable event (e.g., "The negotiations ended in \*an atmosphere of mutual distrust\*").
4. **\*\*Micro-Events\*\***: Insert a plausible but unrecorded specific physical action into a real event (e.g., "The diplomat \*refused to shake hands\* before the meeting started").
5. **\*\*Unfolding Details\*\***: Add unverifiable specifics about the timing, location, or participants of a verifiable event (e.g., "The CEO \*arrived 15 minutes late\* to the conference", "Trump extended a dinner invitation to the Chinese delegation after the meeting", but this detail was not included in the evidence).

**\*\*Goal\*\***: The claim should sound like a confident assertion (real news from reliable source, e.g., BBC, AP) or analysis, not just a rumor or gossip.

**\*\*Output Format\*\* (JSON)**:

```
{  
  "claim_text": "Your claim here"  
}
```

**\*\*Your JSON output\*\***:

Figure 8: Prompt for Ambiguous Claim Generation

set during training—a direct emulation of real-world BDC. We then fine-tuned three representative models (Qwen3-30B-A3B-Instruct-2507, Qwen3-4B-Instruct-2507, and Llama-3.1-8B-Instruct) on this contaminated corpus using Low-Rank Adaptation (LoRA). The specific hyperparameters used for this process are detailed in Table 7.

We subsequently evaluated both the original "clean" models and their contaminated counterparts (marked with \*) on two versions of the benchmark: the Original LiveFact and the Entity Shift LiveFact. The resource consumption for this contamination simulation is provided in Table 8.

## C.2 SSA Implementation Results

The results of the simulation, presented in Table 10, provide compelling evidence for the efficacy of the SSA framework.

First, observe the performance on the Original LiveFact benchmark. The contaminated models (\*) achieve near-perfect accuracy scores (e.g., 99.89% for Qwen3-30B-A3B-Instruct-2507\* at  $T + 3$ ), a massive jump from the original model's 77.00%.

This confirms that the fine-tuning successfully injected the specific knowledge of the benchmark into the models' parameters; they are essentially "reciting" the memorized answers (Ovadia et al., 2024; Lyu et al., 2024; Li et al., 2025b).

However, the critical insight comes from comparing performance on the Entity Shift LiveFact. When tested on the version where entity names are swapped (e.g., "Trump" → "Wannetta"), the performance of the contaminated models drops precipitously. For Qwen3-30B-A3B-Instruct-2507\*, accuracy falls from 99.89% (Original) to 84.77% (Shifted). While still high due to general instruction tuning, this gap ( $\Delta$ ) exposes the model's reliance on specific entity tokens rather than generalized reasoning.

This effect is quantified by the SSA Factor in Table 9. The original, clean models maintain a negligible SSA Factor (e.g., 0.08 for Qwen3-30B-A3B-Instruct-2507), indicating robust, entity-agnostic reasoning. In stark contrast, the contaminated models exhibit massive spikes in their SSA scores (e.g., 5.18 for Qwen3-30B-A3B-Instruct-2507\* and 6.67

Model Name	Training Type	Architecture	Parameters	Size
<i>Qwen Series</i>				
Qwen3-235B-A22B-Instruct-2507	Instruction	MoE	235B (22B Active)	Large
Qwen3-30B-A3B-Instruct-2507	Instruction	MoE	30B (3B Active)	Medium
Qwen3-30B-A3B	Hybrid	MoE	30B (3B Active)	Medium
Qwen3-30B-A3B-Base	Base	MoE	30B (3B Active)	Medium
Qwen3-32B	Hybrid	Dense	32B	Medium
Qwen3-8B	Hybrid	Dense	8B	Small
Qwen3-8B-Base	Base	Dense	8B	Small
Qwen3-4B-Instruct-2507	Instruction	Dense	4B	Compact
<i>Meta Llama</i>				
Llama-3.3-70B-Instruct	Instruction	Dense	70B	Large
Llama-3.1-70B	Base	Dense	70B	Large
Llama-3.1-8B	Base	Dense	8B	Small
Llama-3.1-8B-Instruct	Instruction	Dense	8B	Small
Llama-3.2-3B	Base	Dense	3B	Compact
Llama-3.2-1B	Base	Dense	1B	Compact
<i>DeepSeek</i>				
DeepSeek-V3.1	Hybrid	MoE	671B (37B Active)	Huge
<i>Moonshot AI</i>				
Kimi-K2-Thinking	Reasoning	MoE	1T (32B Active)	Huge
<i>OpenAI (OSS)</i>				
gpt-oss-120b	Reasoning	MoE	120B	Medium
gpt-oss-20b	Reasoning	MoE	20B	Small
<i>OpenAI (GPT Series)</i>				
gpt-5.2-2025-12-11	Instruction	-	-	-
gpt-5.1-2025-11-13	Instruction	-	-	-
gpt-4o-2024-08-06	Instruction	-	-	-
gpt-4o-mini-2024-07-18	Instruction	-	-	-

Table 5: Overview of LLMs used in the experiments, including their training type, architecture, parameter counts, and relative size classification. Since the OpenAI GPT series models are commercial closed-source models, their related information is unavailable.

for Qwen3-4B-Instruct-2507\*). This drastic increase—driven by both the performance drop ( $\Delta$ ) and the high OTR—serves as a clear, quantifiable signal of data contamination. These findings definitively validate that the SSA framework acts as a sensitive "litmus test" for label-level contamination, justifying its integral role in the long-term maintenance of the LiveFact benchmark.

## D Supplementary Analysis

### D.1 Impact of Training Mode: Base, Instruct, and Hybrid

To isolate the effect of training methodology on reasoning capabilities, we conducted an ablation study comparing different variants of the same model architectures. We examined Base models (pre-trained only), Instruct models (SFT + RLHF), and Hybrid

models across the Qwen3 and Llama-3.1 families. The performance results are detailed in Table 11.

The analysis reveals a stark "Alignment Tax" on performance for models that lack instruction tuning. The Llama-3.1-8B base model collapses to near-random performance (Avg: 5.14%), primarily because it fails to adhere to the strict output formatting required by the benchmark ([[LABEL]]). Without SFT, the model often continues generating context or hallucinates unrelated text instead of providing a classification. In contrast, the Instruct version reaches 49.96%, demonstrating that the knowledge exists within the parameters but requires alignment to be accessible. A similar, though less catastrophic, trend is seen in Qwen3-30B-A3B, where the specific hybrid variant underperformed significantly (1.30%) likely due to a mismatch in prompt template expectations com-

Evaluation Prompt
<p><b>SYSTEM_PROMPT</b> = ""You are a fact-checking expert. Classify news claims based on evidence. Labels:</p> <ul style="list-style-type: none"> <li>- REAL: Claim is supported by evidence, factually accurate</li> <li>- FAKE: Claim contradicts evidence or contains false information</li> <li>- AMBIGUOUS: Claim cannot be definitively verified</li> </ul> <p>IMPORTANT: Always start your response with the classification label in double brackets. Format: [[LABEL]] followed by brief reasoning. Example response: [[REAL]] The claim is supported by multiple news sources reporting the same event details.""</p> <p><b>USER_PROMPT_TEMPLATE</b> = ""Context: {context} Claim: "{claim}" Evidence: {evidence} Classify this claim. Start with [[REAL]], [[FAKE]], or [[AMBIGUOUS]], do not think too much.""</p>

Figure 9: Prompt for Evaluation

pared to the robust Instruct-2507 version (69.46%).

However, an interesting anomaly emerges with Qwen3-8B-Base. Unlike its Llama counterpart, this base model exhibits surprisingly strong reasoning capabilities, achieving an average score of 61.10%, which is competitive with the Qwen3-8B hybrid model (63.62%). Notably, in the  $\delta = -3$  Inference Mode, the Base model actually outperforms the hybrid version (83.08% vs 60.29%). This suggests that the Qwen3 base pre-training corpus may contain richer, more structured reasoning patterns that allow it to naturally infer the task even without explicit instruction tuning. This finding challenges the assumption that base models are universally unsuitable for zero-shot reasoning tasks, provided the underlying pre-training is sufficiently robust.

## D.2 The Cost of Reasoning: Verbosity vs. Compliance in Thinking Models

A critical finding from our main experiments is the distinct behavior of models heavily optimized for "Reasoning" or "Chain-of-Thought" (CoT), specifically Kimi-K2-Thinking and the GPT-OSS series. In our initial screening using a standard 128-token limit (Table 12), these models appeared to fail catastrophically (e.g., Kimi-K2-Thinking scored ~1.76%). However, upon qualitative inspection (Figure 11), it became evident that the models were not hallucinating but rather engaging in extensive

Model Name	Tokens (M)	Time (h)	Cost (\$)
<i>Qwen Series</i>			
Qwen3-235B-A22B-Instruct-2507	4.86	8.20	3.65
Qwen3-30B-A3B-Instruct-2507	4.86	4.09	0.64
Qwen3-30B-A3B	5.09	2.87	0.71
Qwen3-30B-A3B-Base	4.72	1.86	0.60
Qwen3-32B	4.87	5.38	2.72
Qwen3-8B	4.88	2.37	0.73
Qwen3-8B-Base	4.73	2.36	0.67
Qwen3-4B-Instruct-2507	4.96	3.13	0.42
<i>DeepSeek</i>			
DeepSeek-V3.1	4.35	3.53	5.3
<i>Meta Llama</i>			
Llama-3.3-70B-Instruct	4.54	2.04	4.77
Llama-3.1-70B	4.36	3.02	5.22
Llama-3.1-8B	4.40	2.07	0.67
Llama-3.1-8B-Instruct	4.42	2.10	0.63
Llama-3.2-3B	4.37	1.71	0.30
Llama-3.2-1B	4.61	1.74	0.14
<i>Moonshot AI</i>			
Kimi-K2-Thinking	4.71	14.88	5.43
Kimi-K2-Thinking*	7.62	60.92	12.53
<i>OpenAI (OSS)</i>			
gpt-oss-120b	4.06	4.46	1.03
gpt-oss-120b*	5.50	7.00	1.28
gpt-oss-20b	4.92	2.56	0.69
gpt-oss-20b*	5.84	5.71	0.97
<i>OpenAI (GPT Series)</i>			
gpt-5.2-2025-12-11	2.67	2.41	9.27
gpt-5.1-2025-11-13	2.66	1.93	6.44
gpt-4o-2024-08-06	2.30	1.37	2.28
gpt-4o-mini-2024-07-18	2.48	1.69	0.79

Table 6: Resource consumption analysis for each model on the benchmark dataset. Metrics include total tokens processed, total wall-clock time in hours, and API cost in USD. Results averaged from three runs ( $T - 3$ ,  $T$ ,  $T + 3$ ).

internal deliberation.

As shown in Table 12, re-evaluating these models with an extended 1024-token window unlocked their true potential. gpt-oss-120b, for instance, jumped to an average accuracy of 72.13%, rivaling the best-performing Qwen models. This reveals a fundamental trade-off in current model design: "Thinking Models" prioritize exhaustive reasoning paths over strict formatting constraints. Unlike standard Instruction-Tuned models (e.g., Qwen3-Instruct) that can "snap" to a label immediately, Reasoning models treat the prompt as a starting point for a dialectical process. While this leads to high accuracy when resources are unconstrained (Table 6), it poses significant challenges for low-latency, automated benchmarking environments where conciseness is often a proxy for adherence.

Hyperparameter	Value
Fine-tuning Method	LoRA
LoRA Rank ( $r$ )	8
Dropout	0.05
Learning Rate	$1 \times 10^{-4}$
Optimizer	Adam
Batch Size	4
Number of Epochs	3
Loss Function	Cross-Entropy

Table 7: Fine-tuning hyperparameters for BDC simulation experiments.

Model Name	Tokens (M)	Time (h)	Cost (\$)
Qwen3-30B-A3B-Instruct-2507	24.12	3.69	5.18
Qwen3-4B-Instruct-2507	24.5	2.32	3.24
Llama-3.1-8B-Instruct	22.19	2.32	5.27

Table 8: Resource consumption for the contamination simulation experiments.

### D.3 Deep Dive: Confusion Matrix Analysis

To provide a granular view of model behavior, we analyzed the confusion matrices across all model groups (Figures 12).

Across all valid models (e.g., Qwen3-Instruct, GPT-4o), we observe a massive shift in prediction behavior at  $T - 3$ . In Classification Mode, the matrices show a "spray" of predictions; models struggle to differentiate classes, often defaulting to the majority class of their training data. However, in Inference Mode, the confusion matrix tightens significantly around the "Ambiguous" class. For example, Qwen3-32B correctly concentrates nearly all its mass on "Ambiguous" at  $T - 3$  (Inference), whereas in Classification mode, it erroneously predicts "Real" or "Fake" with low confidence. This confirms that the models detect the ambiguity but are forced to hallucinate in standard settings.

As time progresses to  $T$  and  $T + 3$ , the confusion matrices for high-performing models (GPT-5.2, Qwen3-235B) show a distinct "diagonalization." The noise from the "Ambiguous" class dissipates, and predictions sharpen into correct "Real" vs. "Fake" classifications. This temporal evolution proves that the benchmark successfully captures the arrival of new information.

The confusion matrices for base models (e.g., Llama-3.1-8B) reveal a pathological failure mode. Rather than a distributed error pattern, these models often exhibit a single-column collapse, predicting

Model	$\Delta$	OTR	SSA
Qwen3-30B-A3B-Instruct-2507	4.90	1.53	0.08
Qwen3-30B-A3B-Instruct-2507*	15.24	33.99	5.18
Qwen3-4B-Instruct-2507	6.56	2.76	0.18
Qwen3-4B-Instruct-2507*	14.58	45.72	6.67
Llama-3.1-8B-Instruct	1.46	5.31	0.08
Llama-3.1-8B-Instruct*	13.53	39.16	5.30

Table 9: SSA Factor calculation showing the clear distinction between clean and contaminated (\*) models. The calculation of  $\Delta$  and OTR is based on the average of six runs performed under the selected Classification and Inference modes.

one class (e.g., "Real") for 100% of queries regardless of input. This is not due to reasoning but due to a failure to condition on the evidence; the model likely defaults to the most common token in its pre-training distribution for the given prompt context. This visualizes exactly why instruction tuning is non-negotiable for reliable automated fact-checking.

### D.4 Error Analysis on Frontier Models

While Section 4 provides an aggregate view of model performance across the LiveFact benchmark, it is crucial to diagnose the specific failure modes of SOTA frontier models. Interestingly, our analysis reveals that the primary source of errors for highly capable models is fundamentally different from that of earlier generations. Specifically, we observe a phenomenon we term *Over-Safety* (or extreme *Epistemic Humility*), where models prioritize cautious refusals over definitive verdicts.

To quantify this, we conducted a deep-dive error breakdown on **GPT-5.2** under the Inference Mode at time slice  $T$ . We isolated all claims where the Ground Truth is strictly "Fake" (i.e., the provided evidence slice contains explicit contradictions to the claim). Out of 1,390 such instances in the November 2025 data, GPT-5.2 correctly predicted "Fake" in 83.24% of cases. For the remaining erroneous predictions, the breakdown is as follows:

- **Hallucinations (Predicted "Real"):** Only **0.22%** (3 instances).
- **False Negatives via Ambiguity (Predicted "Ambiguous"):** **10.58%** (147 instances).
- **Safety Refusals (Null Output):** **5.97%** (83 instances).

Model	Eval. Mode	Classification						Inference						Avg.
		Acc <sub>-3</sub> <sup>CLS</sup>	Acc <sub>0</sub> <sup>CLS</sup>	Acc <sub>+3</sub> <sup>CLS</sup>	F1 <sub>-3</sub> <sup>CLS</sup>	F1 <sub>0</sub> <sup>CLS</sup>	F1 <sub>+3</sub> <sup>CLS</sup>	Acc <sub>-3</sub> <sup>INF</sup>	Acc <sub>0</sub> <sup>INF</sup>	Acc <sub>+3</sub> <sup>INF</sup>	F1 <sub>-3</sub> <sup>INF</sup>	F1 <sub>0</sub> <sup>INF</sup>	F1 <sub>+3</sub> <sup>INF</sup>	
<b>Original LiveFact</b>														
Qwen3-30B-A3B-Instruct-2507		55.24	75.05	77.00	52.39	74.87	76.61	64.55	74.64	76.78	55.37	74.66	76.40	69.46
Qwen3-30B-A3B-Instruct-2507*		84.59	99.57	99.89	83.28	99.57	99.89	59.24	97.13	99.52	52.42	97.16	99.52	89.31
Qwen3-4B-Instruct-2507		46.29	62.43	63.39	42.63	60.70	61.58	83.08	62.45	63.39	64.45	61.15	61.65	61.10
Qwen3-4B-Instruct-2507*		77.98	98.98	99.50	74.50	98.97	99.50	53.05	96.63	99.16	52.16	96.66	99.16	87.19
Llama-3.1-8B-Instruct		40.69	50.05	50.20	36.53	45.67	45.11	78.69	50.16	50.09	60.26	46.89	45.14	49.96
Llama-3.1-8B-Instruct*		88.84	96.02	96.49	84.37	96.02	96.49	46.02	93.62	96.17	42.14	93.62	96.17	85.50
<b>Entity Shift LiveFact</b>														
Qwen3-30B-A3B-Instruct-2507		46.93	69.85	71.40	43.17	68.82	70.16	81.63	69.08	71.31	63.35	68.89	70.08	64.56
Qwen3-30B-A3B-Instruct-2507*		69.60	84.31	84.77	66.39	82.61	84.74	50.02	76.25	86.34	41.16	76.35	86.25	74.07
Qwen3-4B-Instruct-2507		44.22	61.18	63.91	37.98	60.49	62.95	36.16	60.61	63.87	39.56	60.61	62.99	54.54
Qwen3-4B-Instruct-2507*		67.33	81.90	83.49	62.97	82.09	83.50	47.75	75.96	85.31	39.70	76.03	85.26	72.61
Llama-3.1-8B-Instruct		39.89	48.27	48.54	34.87	43.16	42.96	79.67	48.38	48.47	60.32	44.42	43.05	48.50
Llama-3.1-8B-Instruct*		71.43	79.62	82.67	69.57	79.72	82.70	45.47	73.54	83.49	38.29	73.72	83.36	71.97

Table 10: Performance comparison of original vs. contaminated (\*) models on Original and Entity-Shifted LiveFact datasets.

Model	Eval. Mode	Classification						Inference						Avg.
		Acc <sub>-3</sub> <sup>CLS</sup>	Acc <sub>0</sub> <sup>CLS</sup>	Acc <sub>+3</sub> <sup>CLS</sup>	F1 <sub>-3</sub> <sup>CLS</sup>	F1 <sub>0</sub> <sup>CLS</sup>	F1 <sub>+3</sub> <sup>CLS</sup>	Acc <sub>-3</sub> <sup>INF</sup>	Acc <sub>0</sub> <sup>INF</sup>	Acc <sub>+3</sub> <sup>INF</sup>	F1 <sub>-3</sub> <sup>INF</sup>	F1 <sub>0</sub> <sup>INF</sup>	F1 <sub>+3</sub> <sup>INF</sup>	
Qwen3-30B-A3B-Instruct-2507		55.24	75.05	77.00	52.39	74.87	76.61	64.55	74.64	76.78	55.37	74.66	76.40	69.46
Qwen3-30B-A3B		3.42	0.14	0.07	6.05	0.27	0.14	2.69	0.11	0.07	2.27	0.23	0.14	1.30
Qwen3-30B-A3B-Base		46.47	55.17	54.92	43.26	51.67	51.33	76.34	54.58	54.78	55.80	51.57	51.25	53.93
Qwen3-8B		50.84	68.88	69.83	49.12	67.79	68.44	60.29	68.67	69.65	53.76	67.82	68.30	63.62
Qwen3-8B-Base		46.29	62.43	63.39	42.63	60.70	61.58	83.08	62.45	63.39	64.45	61.15	61.65	61.10
Llama-3.1-8B-Instruct		40.69	50.05	50.20	36.53	45.67	45.11	78.69	50.16	50.09	60.26	46.89	45.14	49.96
Llama-3.1-8B		11.89	3.62	3.12	12.39	5.72	4.99	1.25	3.30	3.12	1.90	5.39	5.02	5.14

Table 11: Ablation study on training methodologies across Qwen and Llama architectures. Hybrid and Instruct models generally outperform Base models, though strong base reasoning is observed in Qwen3-8B.

This striking contrast—a 16.55% false-negative rate driven by caution versus a near-zero (0.22%) hallucination rate—demonstrates that the challenge in AI-driven verification has shifted. For frontier models, the "Reasoning Gap" is less about the incapability to discern the truth, and more about the strict calibration of alignment filters and refusal thresholds when processing sensitive geopolitical news.

To further illustrate this behavioral shift, we present two representative qualitative failure cases in Table 13, highlighting the nuances of *Over-Caution* and *Safety Refusal*.

These findings offer a critical take-home message for the community: as LLMs become more aligned, static benchmarks that only measure binary accuracy ("Real" vs. "Fake") fail to capture the "alignment tax" imposed on verification tasks. LiveFact’s inclusion of the "Ambiguous" class and dynamic slices effectively exposes this nuance, set-

ting a more precise standard for future model evaluation.

## D.5 Longitudinal Evaluation: December 2025 Version

To demonstrate the temporal consistency of model behaviors and validate the long-term sustainability of the LiveFact framework—a critical requirement for combating BDC—we conducted a longitudinal evaluation using the newly collected December 2025 data snapshot. We evaluated a representative subset of 8 hybrid/instruction-tuned open-source models from our main experiments.

As shown in Table 14, the December 2025 dataset contains 4,222 claims, which are perfectly balanced across the "Real" (1,404), "Fake" (1,406), and "Ambiguous" (1,412) labels in the time-invariant Classification Mode. Crucially, the dynamic label distribution in Inference Mode exhibits the exact same "fog of war" pattern observed

Model	Eval. Mode	Classification						Inference						Avg.
		Acc <sub>-3</sub> <sup>CLS</sup>	Acc <sub>0</sub> <sup>CLS</sup>	Acc <sub>+3</sub> <sup>CLS</sup>	F1 <sub>-3</sub> <sup>CLS</sup>	F1 <sub>0</sub> <sup>CLS</sup>	F1 <sub>+3</sub> <sup>CLS</sup>	Acc <sub>-3</sub> <sup>INF</sup>	Acc <sub>0</sub> <sup>INF</sup>	Acc <sub>+3</sub> <sup>INF</sup>	F1 <sub>-3</sub> <sup>INF</sup>	F1 <sub>0</sub> <sup>INF</sup>	F1 <sub>+3</sub> <sup>INF</sup>	
Kimi-K2-Thinking		6.72	0.20	0.23	10.51	0.40	0.45	0.64	0.16	0.23	0.82	0.33	0.45	1.76
Kimi-K2-Thinking*		45.97	57.15	54.21	45.34	61.71	59.22	58.13	56.56	54.08	42.07	61.57	59.15	54.60
gpt-oss-120b		27.19	20.13	19.54	33.87	30.98	29.85	26.48	19.67	19.51	25.04	31.00	29.90	26.10
gpt-oss-120b*		55.83	79.94	81.81	52.82	79.79	81.49	62.23	78.89	81.60	50.83	78.97	81.31	72.13
gpt-oss-20b		18.76	23.63	23.34	26.02	33.39	32.80	16.44	23.34	23.32	22.07	33.86	32.88	25.82
gpt-oss-20b*		47.84	65.55	67.42	47.16	61.33	62.32	41.46	64.32	67.28	36.53	61.07	62.34	57.05

Table 12: Performance of Reasoning Models evaluated with extended token limits (1024 tokens). The asterisk (\*) denotes the non-standard evaluation setting required to accommodate verbose Chain-of-Thought outputs.

Failure Type 1: Over-Caution (Ground Truth: FAKE → Prediction: AMBIGUOUS)	
<b>Claim ID</b>	EV20251101-10-4 (Inference Mode with $E^{(0)}$ )
<b>Context</b>	A specific allegation regarding a senior US diplomat privately demanding Israel to free exactly 200 Palestinian security prisoners during White House talks.
<b>Ground Truth</b>	FAKE (The evidence explicitly confirms no such specific demand or exact number was discussed, contradicting the claim).
<b>GPT-5.2 Output</b>	[[[AMBIGUOUS]]] The cited coverage focuses on negotiations and hostage/body returns during a Gaza truce, but it does not clearly and consistently corroborate the specific allegation that a ‘senior US diplomat’ privately demanded Israel free exactly 200 Palestinian security prisoners in White House talks...
<b>Analysis</b>	The model successfully identifies that the broad topic (Gaza truce) is present in the evidence but the specific fabricated details are missing. However, instead of confidently deducing that the claim is FAKE due to these explicit contradictions with the official minutes provided in the slice, it conservatively retreats to AMBIGUOUS, demanding an impossible level of direct negation.
Failure Type 2: Safety Refusal (Ground Truth: FAKE → Prediction: Null/Refusal)	
<b>Claim ID</b>	EV20251101-20-4 (Inference Mode with $E^{(0)}$ )
<b>Context</b>	A fabricated claim involving sensitive military actions and political figures in a conflict zone.
<b>Ground Truth</b>	FAKE
<b>GPT-5.2 Output</b>	"" (Empty string / System refusal)
<b>Analysis</b>	Despite being provided with a golden evidence slice that disproves the claim, the model’s internal safety guardrails are triggered by the sensitive keywords (e.g., specific conflict zones, names of political leaders). It halts the generation process entirely, leading to a missing prediction.

Table 13: Representative failure cases for GPT-5.2 demonstrating Over-Safety and Epistemic Humility. Even when sufficient evidence is present to declare a claim "Fake", frontier models often default to "Ambiguous" or trigger safety refusals.

in the November 2025 data. At the  $E^{(-3)}$  slice, the vast majority of claims (3,533 out of 4,222) shift to the "Ambiguous" class, accurately reflecting the real-world scarcity of preemptive evidence before an event unfolds. As time progresses to  $E^{(0)}$  and  $E^{(+3)}$ , the distribution naturally aligns back with the ground-truth Classification labels as evidence becomes sufficient.

Evidence Set	Real	Fake	Ambiguous	Total Claims
$E^{(-3)}$	343	346	3,533	4,222
$E^{(0)}$	1,374	1,376	1,472	4,222
$E^{(+3)}$	1,404	1,406	1,412	4,222

Table 14: Dynamic Label Distribution for Inference Mode (December 2025)

The evaluation results on the December 2025 version (Table 15) highly correlate with our find-

ings from November 2025, confirming the temporal robustness of the benchmark. The Qwen3-235B-A22B-Instruct model consistently maintains its leading position across almost all metrics, achieving the highest overall average score (71.02%), followed closely by the Qwen3-30B-A3B model.

Furthermore, the December results strongly corroborate our observations regarding the *Reasoning Gap* and *Epistemic Humility*. For instance, models like DeepSeek-V3.1 and Qwen3-32B exhibit relatively low accuracy in the Classification Mode at  $E^{(-3)}$  (43.46% and 45.81%, respectively), which primarily measures prior bias or forced guessing. However, when evaluated in Inference Mode at the same  $E^{(-3)}$  slice—where they are allowed to assess evidence sufficiency—their accuracy surges to 82.76% and 81.08%. This massive leap demonstrates a consistent capacity for epistemic humility

Eval. Mode Model	Classification						Inference						Avg.
	Acc <sub>-3</sub> <sup>CLS</sup>	Acc <sub>0</sub> <sup>CLS</sup>	Acc <sub>+3</sub> <sup>CLS</sup>	F1 <sub>-3</sub> <sup>CLS</sup>	F1 <sub>0</sub> <sup>CLS</sup>	F1 <sub>+3</sub> <sup>CLS</sup>	Acc <sub>-3</sub> <sup>INF</sup>	Acc <sub>0</sub> <sup>INF</sup>	Acc <sub>+3</sub> <sup>INF</sup>	F1 <sub>-3</sub> <sup>INF</sup>	F1 <sub>0</sub> <sup>INF</sup>	F1 <sub>+3</sub> <sup>INF</sup>	
Qwen3-235B-A22B-Instruct-2507	<u>52.18</u>	<b>78.83</b>	<b>80.53</b>	47.97	<b>78.93</b>	<b>80.57</b>	62.65	<b>78.56</b>	<b>80.53</b>	52.25	<b>78.72</b>	<b>80.57</b>	<b>71.02</b>
Qwen3-30B-A3B-Instruct-2507	<b>55.38</b>	<u>73.80</u>	<u>75.49</u>	<b>52.35</b>	<u>73.35</u>	<u>74.90</u>	63.60	<u>73.26</u>	<u>75.49</u>	54.35	<u>72.92</u>	<u>74.90</u>	<u>68.32</u>
Qwen3-32B	45.81	70.72	71.06	40.57	69.14	69.14	<u>81.08</u>	70.91	71.06	<u>60.09</u>	69.28	69.14	61.85
Qwen3-8B	50.85	67.22	67.10	<u>48.65</u>	65.36	64.72	54.95	66.94	67.10	50.88	65.30	64.72	61.15
Qwen3-4B-Instruct-2507	45.31	63.31	66.49	37.59	62.52	65.61	35.03	62.96	66.49	36.30	62.50	65.61	55.81
Llama-3.3-70B-Instruct	44.91	68.31	68.95	38.59	62.23	62.34	22.55	67.10	68.95	31.97	61.38	62.34	54.97
Llama-3.1-8B-Instruct	41.76	49.53	49.55	37.88	44.96	44.68	79.06	49.74	49.55	<b>61.26</b>	45.74	44.68	49.87
DeepSeek-V3.1	43.46	72.64	74.49	36.01	70.81	72.67	<b>82.76</b>	72.57	74.49	57.83	70.68	72.67	66.76

Table 15: Performance comparison of 8 representative LLMs on LiveFact using the December 2025 longitudinal data. The best result in each column is **bolded**, and the second-best result is underlined.

across different time periods, as these models successfully recognize missing evidence and output the "Ambiguous" label.

Conversely, Llama-3.3-70B-Instruct struggles significantly at Inference  $E^{(-3)}$  (22.55%), indicating a persistent behavioral flaw where it overcommits to binary verification even when evidence is objectively insufficient.

In summary, the successful deployment of the December 2025 evaluation not only proves the logistical sustainability of LiveFact’s monthly update pipeline but also verifies that the model capabilities and failure modes identified in our primary November study are consistent and generalizable over time.

### Entity Shift Prompt

**SYSTEM\_PROMPT** = ""You are a precise data processor for research. Your task is to perform entity shifting - replacing real-world named entities with fictional alternatives while preserving all semantic meaning.

You must:

1. Replace ALL named entities (people, organizations, cities, locations, nationalities, races) with fictional alternatives, ensure these names are not widely recognized
2. Use simple, uncommon but plausible names (e.g., "Trump" → "Korwin", "United States" → "Northland", "BBC" → "GBN")
3. Keep entity types consistent (person→person, country→country, organization→organization)
4. Preserve ALL other information (dates, numbers, events, relationships, facts)
5. Ensure the SAME entity always maps to the SAME fictional name across ALL fields
6. Output valid JSON only, no markdown code blocks, no explanations""

**ENTITY\_SHIFT\_PROMPT** = ""Perform entity shifting on the following news data. Replace all named entities with fictional alternatives while preserving semantic meaning.

**\*\*Original Event Title\*\***: {event\_title}

**\*\*Original Context\*\***: {context}

**\*\*Original Claim\*\***: {claim}

**\*\*Original Evidence Titles\*\*** (shift ONLY these titles, keep source names as-is):

{evidence\_titles}

**\*\*Requirements\*\***:

1. Identify ALL named entities: PERSON names, ORGANIZATION names, LOCATION names, NATIONALITY terms
2. Replace each with a LESS RECOGNIZABLE fictional alternative
3. Use simple, uncommon but plausible names (e.g., "Prince Andrew" → "Lord Harwick", "Buckingham Palace" → "Thornfield Manor", "United Kingdom" → "Alberia")
4. Keep entity types consistent (person→person, palace→palace, country→country)
5. PRESERVE all roles, positions, relationships, dates, numbers, and facts - ONLY change entity names
6. Ensure CONSISTENT mapping: the SAME original entity ALWAYS maps to the SAME fictional name in ALL fields
7. Do NOT change news source names (BBC, Reuters, etc.) - these are metadata, not content entities
8. Evidence titles should use the SAME entity mapping as claim, context, and event\_title

**\*\*Output Format\*\*** (JSON only, NO markdown, NO code blocks, NO explanations):

```
{{
  "entity_mapping": {{
    "Original Entity 1": "Fictional Name 1",
    "Original Entity 2": "Fictional Name 2"
  }},
  "event_title_shifted": "shifted event title here",
  "context_shifted": "shifted context here",
  "claim_shifted": "shifted claim here",
  "evidence_titles_shifted": [
    "shifted evidence title 1",
    "shifted evidence title 2"
  ]
}}
```

Output JSON only:""

Figure 10: Prompt for Entity Shift Processing

### Case Study: Kimi-K2-Thinking vs. Qwen3-235B-A22B-Instruct-2507

**Evaluator:** You are a fact-checking expert. Classify news claims based on evidence.....Claim: One student escaped as authorities intensified their search for the 24 other girls abducted by gunmen from Government Girls College in Maga, Kebbi state, northwestern Nigeria.....Classify this claim. Start with `[[REAL]]`, `[[FAKE]]`, or `[[AMBIGUOUS]]`, do not think too much.

**Kimi-K2-Thinking:** <think> The user wants me to classify a news claim based on the provided evidence. The claim is:\n\n\"One student escaped as authorities intensified their search for the 24 other girls abducted by gunmen from Government Girls College in Maga, Kebbi state, northwestern Nigeria.\"\n\nLet me analyze the evidence:\n\nKey elements of the claim:\n1. Location: Government Girls College in Maga, Kebbi state, northwestern Nigeria\n2. Event: Gunmen abducted girls\n3. Number: 24 other girls (implying 25 total, with 1 escaped)\n4. One student escaped\n5

**Qwen3-235B-A22B-Instruct-2507:** `[[REAL]]` Multiple credible sources confirm that 25 girls were abducted from Government Girls College Maga in Kebbi State, and at least one student escaped, with authorities intensifying searches for the others. Sources like BBC, CNN, NPR, and Al Jazeera consistently report these details.

Figure 11: Comparison of raw model outputs. Kimi-K2 fails to produce the required classification label, getting stuck in an internal "thought chain," whereas Qwen3 immediately delivers the structured verdict.

### Confusion Matrices: Classification (CLS) vs Inference (INF) (Part 1/4)

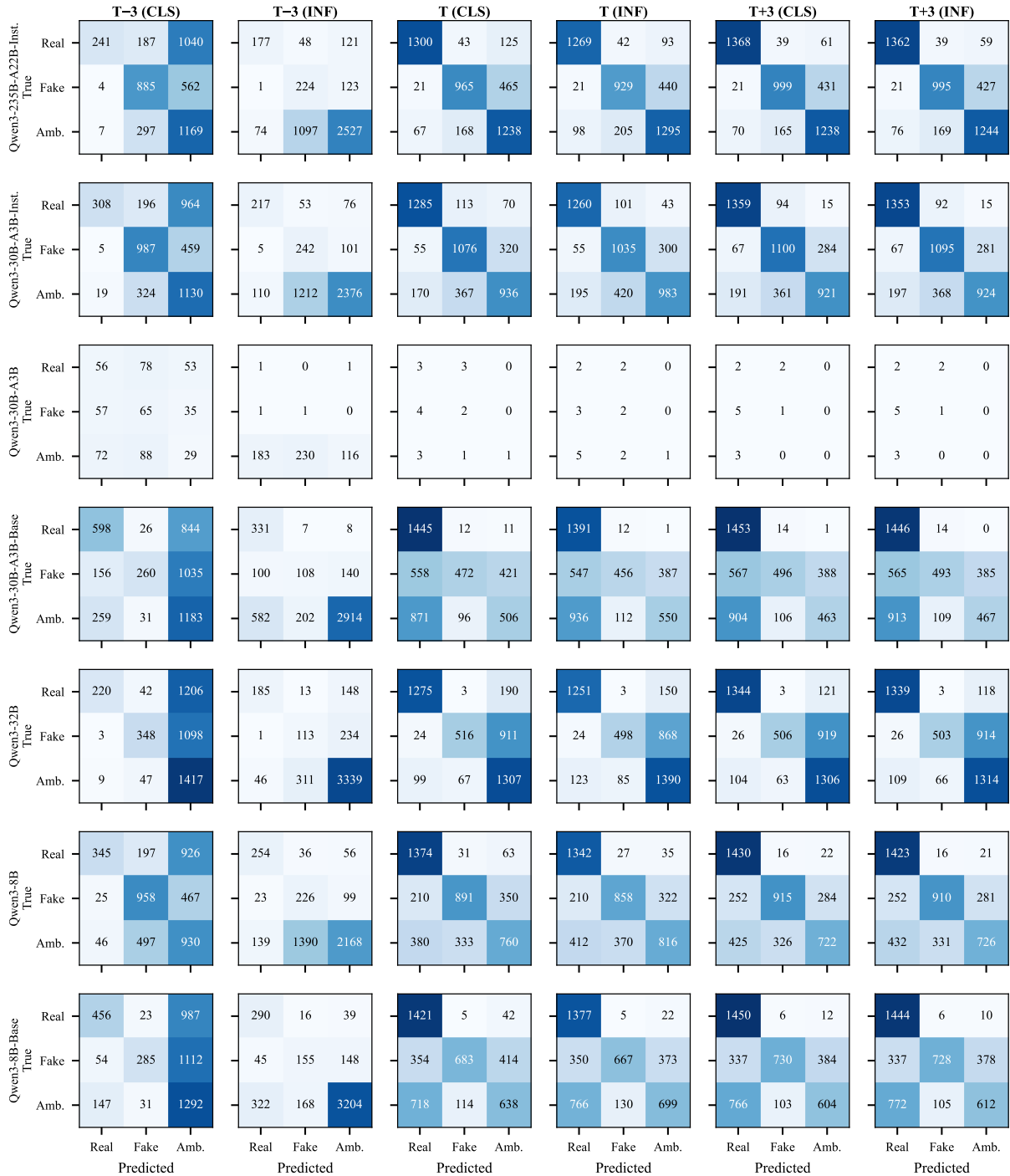


Figure 12: Confusion Matrices (Part 1/4) - Qwen Series. Note the sharp contrast in diagonal clarity between Instruct models and Base models.

Confusion Matrices: Classification (CLS) vs Inference (INF) (Part 2/4)

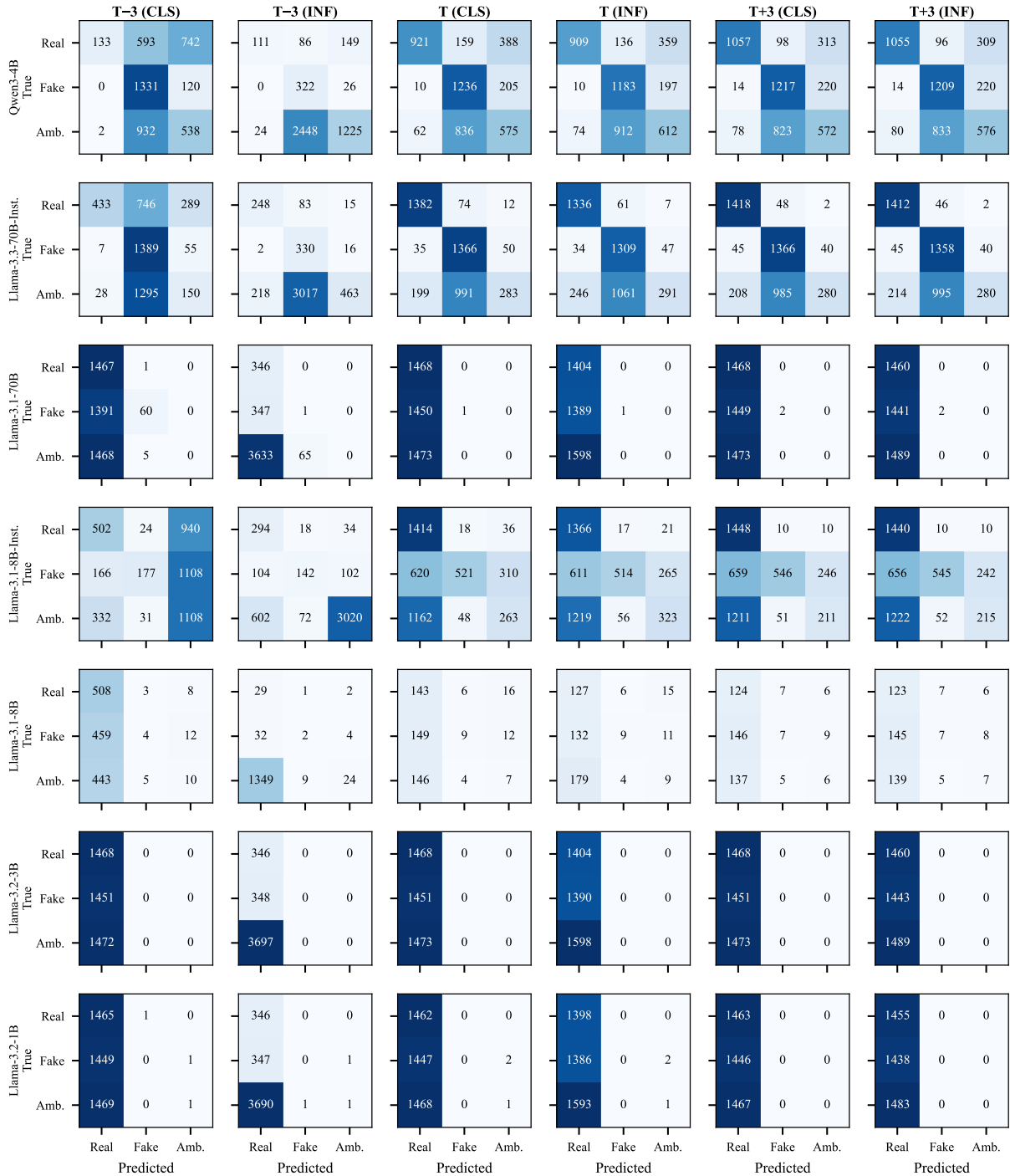


Figure 12 (cont.): Confusion Matrices (Part 2/4) - Qwen, Llama Series. Llama-3.1-70B, Llama-3.1-8B, Llama-3.2-3B, and Llama-3.2-1B exhibit "collapsing" behavior, predicting a single class.

### Confusion Matrices: Classification (CLS) vs Inference (INF) (Part 3/4)

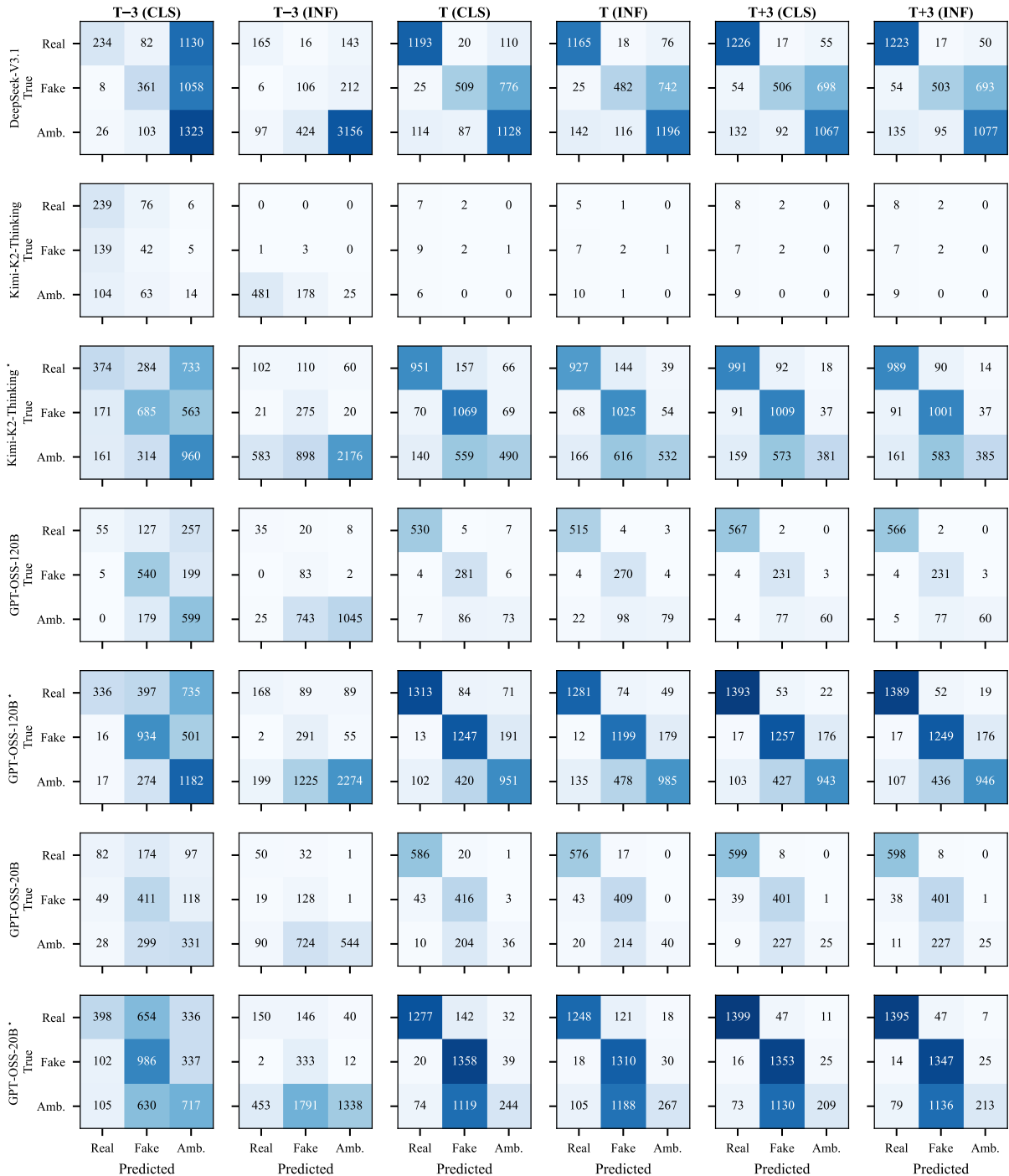


Figure 12 (cont.): Confusion Matrices (Part 3/4) - DeepSeek, Kimi, and GPT-OSS Series. The asterisk \* denotes the non-standard evaluation setting required to accommodate verbose CoT outputs.

**Confusion Matrices: Classification (CLS) vs Inference (INF) (Part 4/4)**

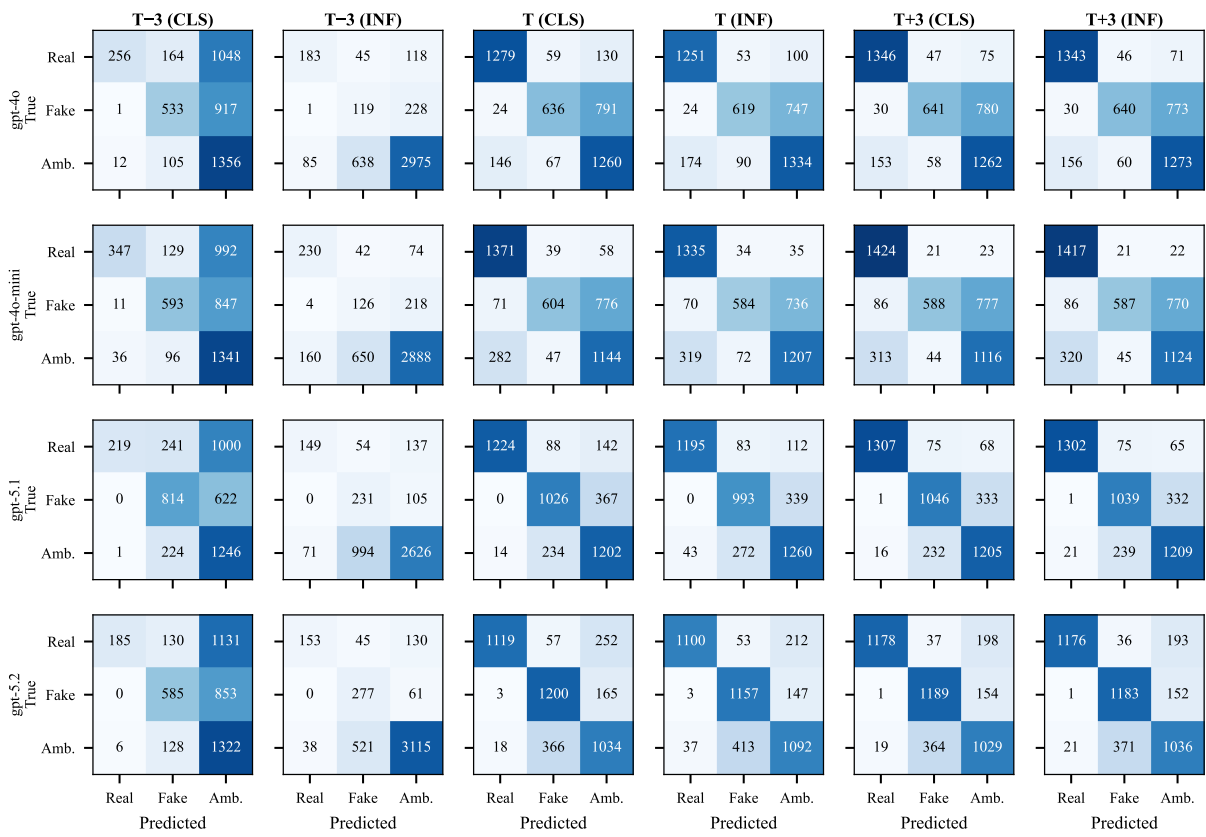


Figure 12 (cont.): Confusion Matrices (Part 4/4) - GPT Series.