

All That Glitters Is Not Gold: A Benchmark for Reference-Free Counterfactual Financial Misinformation Detection

Yuechen Jiang¹, Zhiwei Liu^{1*}, Yupeng Cao², Yueru He³, Ziyang Xu⁴, Chen Xu⁴, Zhiyang Deng², Prayag Tiwari⁵, Xi Chen⁶, Alejandro Lopez-Lira⁷, Jimin Huang^{1,4}, Junichi Tsujii⁸, Sophia Ananiadou^{1,9}

¹University of Manchester, ²Stevens Institute of Technology, ³Columbia University,

⁴The Fin AI, ⁵Halmstad University, ⁶New York University, ⁷University of Florida,

⁸National Institute of Advanced Industrial Science and Technology ⁹ELLIS Manchester

*Corresponding author: zhiwei.liu@manchester.ac.uk

Abstract

We introduce **RFC-BENCH**, a benchmark for evaluating large language models on financial misinformation under realistic news. **RFC-BENCH** operates at the paragraph level and captures the contextual complexity of financial news where meaning emerges from dispersed cues. The benchmark defines two complementary tasks: reference-free misinformation detection and comparison-based diagnosis using paired original–perturbed inputs. Experiments reveal a consistent pattern: performance is substantially stronger when comparative context is available, while reference-free settings expose significant weaknesses, including unstable predictions and elevated invalid outputs. These results indicate that current models struggle to maintain coherent belief states without external grounding. By highlighting this gap, **RFC-BENCH** provides a structured testbed for studying reference-free reasoning and advancing more reliable financial misinformation detection in real-world settings. *we release the dataset at: <https://github.com/lzw108/FMD>.*

1 Introduction

Large Language Models (LLMs) are commonly evaluated on how accurately they interpret fluent text, but they are rarely assessed on whether the text itself is admissible as an object of interpretation (Greshake et al., 2023; Tang et al., 2025; Yu et al., 2025a). From a pragmatic perspective, surface plausibility is not the primary object of interest; instead, what matters is the set of warranted assertions a paragraph puts “on the table” for belief revision under a conversational or decision context (cf. Stalnaker’s theory of common ground (Stalnaker, 2002)). In financial text, minimal edits can maintain fluency while substantially shifting these commitments, for example by turning possibility into certainty or by turning temporal sequence into causation, yielding a counterfac-

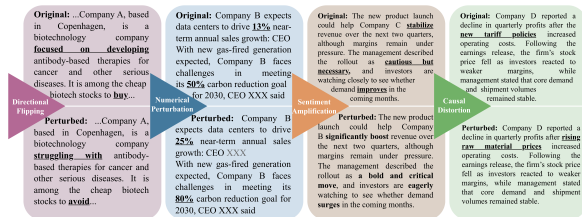


Figure 1: Counterfactual financial misinformation generated via minimal yet belief-shifting edits.

tual world that reads smoothly and adds no new verifiable fact (Figure 1) (Rangapur et al., 2023b; Liu et al., 2025b; Rangapur et al., 2025). Such perturbations often exploit language features that humans can flag as subtly misleading based on discourse-internal cues alone, especially for domain experts and frequently even for careful non-experts (Kahan et al., 2017; Ramos and Van Boven, 2025). It remains unclear whether LLMs show the same counterfactual awareness when given only the perturbed paragraph, with no original version and no external verification. **Will they notice that something is off, or will they accommodate it and produce a confident takeaway?** This makes counterfactual perturbations a practical attack surface as LLMs proliferate in financial applications (Nie et al., 2024; Fu, 2025; Securities and Authority, 2025) amid pervasive financial misinformation (Rangapur et al., 2023b).

Existing misinformation benchmarks largely assume access to external evidence or retrieval mechanisms and thus focus on validating claims with supporting or refuting documentation rather than detecting internal shifts in what a text warrants. For instance, GROVER frames the problem as article-level authenticity detection (Zellers et al., 2019), while FEVER and SCIFACT cast it as claim-level verification with supporting evidence (Thorne et al., 2018; Wadden et al., 2020). Recent financial-domain benchmarks largely inherit the same contract. FIN-FACT annotates claim veracity with

Dataset	Domain	Text Granularity	Flipping	Numerical	Sentiment	Causal	Human/Expert
GROVER	General	Article Level	✗	✗	✗	✗	✗
FEVER	General	Claim Level	✗	✗	✗	✗	✓
SCIFACT	Biomedical	Claim Level	✗	✗	✗	✗	✓
SCITAB	Scientific table	Claim Level	●	●	✗	✗	✓
ContractNLI	Law	Claim/Hypothesis	✗	✗	✗	✗	✓
Fin-Fact	Finance	Claim Level	✗	✗	✗	✗	✓
FINDVER	Finance	Claim Level	●	●	✗	✗	✓
FISCAL	Finance	Claim level	✓	✗	✗	✗	✗
RFC-BENCH (ours)	Finance	Paragraph-level	✓	✓	✓	✓	✓

Table 1: Comparison of misinformation datasets across domains, text granularity, and manipulation dimensions. The table contrasts existing benchmarks with **RFC-BENCH** in terms of input domain, text granularity, supported manipulation types (Flipping, Numerical, Sentiment, and Causal), and the availability of human or expert annotation. Symbols denote the level of support: ✓ indicates full support, ✗ indicates the absence of support, and ● denotes partial or limited support.

evidence (Rangapur et al., 2023a), FINDVER evaluates entailment under long-context or retrieval-based settings (Zhao et al., 2024), and FISCAL trains verifiers over claim–document pairs (Sharma et al., 2025). Benchmark scores are tightly coupled to evidence access and retrieval behavior, a coupling flagged as a threat to evaluation validity in recent guidance (Thibault et al., 2025).

To address this gap, we propose **RFC-BENCH**, a benchmark for paragraph-level, reference-free financial misinformation detection. **RFC-BENCH** contains 1826 original–perturbed paragraph pairs drawn from 1826 real-world financial news sources, constructed to preserve surface plausibility while shifting what the paragraph warrants. Following common misinformation patterns summarized in prior surveys (Rangapur et al., 2023b), we operationalize four manipulation categories. **Directional Flipping** reverses the direction of a claim, **Numerical Perturbation** nudges salient quantities, **Sentiment Amplification** strengthens stance toward bullish or bearish interpretations, and **Causal Distortion** recasts sequence or correlation as causation. Perturbations are generated via category-specific LLM-controlled rewriting and retained only if they satisfy automatic minimality constraints and domain-expert validation (category correctness agreement 98.9%; rewrite validity agreement: 93.7%). The paired design enables **Reference-free Detection**, which flags manipulation from a single paragraph with no external grounding, and **Comparative Diagnosis**, which uses the paired original to identify the manipulation type. We report accuracy and macro-F1 (Yang, 1999) with per-type breakdowns, and additionally AUROC (Bradley, 1997) when class balance deviates.

We evaluate 15 open- and closed-source LLMs and observe a sharp asymmetry between recog-

nizing a problematic paragraph and explaining it after the fact. On **Reference-free Detection**, where the model sees only one paragraph and must decide whether it is admissible to accommodate at face value, accuracy stays near chance, peaking at **53.6%** (DeepSeek-reasoner), with **GPT-4.1** at **52.7%**. Under few-shot settings, performance improves only marginally, with the best observed result reaching **56.7%** (LLaMA-3.3-70B, 8-shot). On **Comparative Diagnosis**, where an original paragraph is provided alongside its perturbed counterpart, and surface-level cues are controlled for, accuracy rises to **0.85–0.97**. This gap is consistent with an accommodation-first default. Current LLMs can often localize what changed once an explicit alternative is supplied, but they do not reliably trigger doubt when only a locally plausible, counterfactually perturbed paragraph is available. In pragmatic terms, the models struggle to guard the common ground against commitment-shifting distortions, which is precisely what makes financial misinformation actionable before any external grounding or verification is possible. While the reference-free detection setting (Task 1) reflects real-world scenarios where no external verification is immediately available, the comparative diagnosis setting (Task 2) is designed as a controlled diagnostic probe. It isolates the model’s ability to recognize subtle semantic shifts when explicit alternatives are provided, enabling fine-grained analysis of model behavior beyond end-task performance. Our contributions are summarized as follows:

- We propose **RFC-BENCH**, a benchmark that operationalizes paragraph-level financial misinformation as *plausibility-preserving counterfactual perturbations* that shift what a paragraph warrants, enabling evaluation without external evidence.

- We define two complementary evaluations including **Reference-free Detection** on single paragraphs and **Comparative Diagnosis** on original–perturbed pairs across four manipulation types (directional, numerical, sentiment, causal).
- We benchmark 15 open- and closed-source LLMs and identify a pragmatic bottleneck that models perform well when an explicit comparison is provided, yet remain unreliable at flagging manipulated paragraphs in isolation, consistent with an accommodation-first failure mode.

2 RFC-BENCH

RFC-BENCH is a paragraph-level benchmark for evaluating large language models or Reference-Free Counterfactual Financial Misinformation Detection, constructed from real news articles and their minimally perturbed variants. As illustrated in Figure 2, it proposed a structured pipeline of data collection, perturbation, and annotation, and supports two complementary evaluation tasks with and without external contextual support.

2.1 Task Formulation

We define two evaluation tasks in **RFC-BENCH** to study complementary aspects of LLM robustness to financial misinformation. Let $N \in \mathcal{N}$ denote a financial news paragraph which is either factual or manipulated, where \mathcal{N} is the set of all news paragraphs.

Task 1: Reference-free Detection. This task evaluates whether an LLM can identify financial misinformation from a single document without access to paired references. The model predicts a binary label

$$\mathcal{Y} := \{\text{True}, \text{False}\},$$

where “True” indicates a factual paragraph and “False” indicates a paragraph containing misinformation, based solely on the input document P :

$$y^* = \arg \max_{y \in \mathcal{Y}} P_{\text{LLM}}(y | N). \quad (1)$$

Task 2: Comparative Diagnosis. This task evaluates whether an LLM can recognize the underlying manipulation mechanism when given a side-by-side comparison between a factual paragraph and its minimally perturbed misinformation variant. Each instance consists of a paired input

$(N^{\text{fact}}, N^{\text{mis}})$, where N^{fact} denotes the original factual content and N^{mis} its manipulated counterpart. Let $m \in \mathcal{M}$ denote a manipulation type, where \mathcal{M} is the set of all manipulation types. By comparing the paired documents:

$$m^* = \arg \max_{m \in \mathcal{M}} P_{\text{LLM}}(m | N^{\text{fact}}, N^{\text{mis}}). \quad (2)$$

2.2 Data Curation

Based on the task definitions, we curate **RFC-BENCH** using original–perturbed paragraph pairs from financial news, where perturbations are minimal yet sufficient to instantiate predefined misinformation categories.

2.2.1 Data Acquisition

We collect 1,404 unique financial news articles, each consisting of a title and a summary, from Yahoo Finance¹. The dataset covers 223 publicly traded stocks **B** and spans the period from April 25, 2025 to December 15, 2025. Detailed statistics on the temporal distribution and dataset composition are reported in Appendix E.

2.2.2 Category-specific LLM Rewriting and Prompt Refinement

To enable systematic and interpretable manipulation of financial narratives, we adopt four manipulation categories grounded in common financial misinformation patterns summarized in prior surveys (Rangapur et al., 2023b): **Numerical Perturbation**, **Directional Flipping**, **Sentiment Amplification**, and **Causal Distortion**. A complete mapping from survey-defined categories to the adopted manipulation types is provided in Appendix C. Data categorization follows a two-stage procedure: a rule-based classifier with category-specific keyword patterns is first applied to identify explicit cases (Appendix O), while remaining samples are annotated by GPT-4.1 using a structured prompting scheme (Appendix P).

Each categorized article is then rewritten using GPT-4.1 under carefully designed, category-specific constraints that control the direction and magnitude of semantic distortions. Prompts are iteratively refined through expert validation, and automatic quality control mechanisms are applied to ensure adherence to the intended manipulation constraints. All category-specific rewriting

¹The data are from publicly available pages on [Yahoo Finance](#). No private or restricted information is involved. See Appendix I for details on dataset release and access.

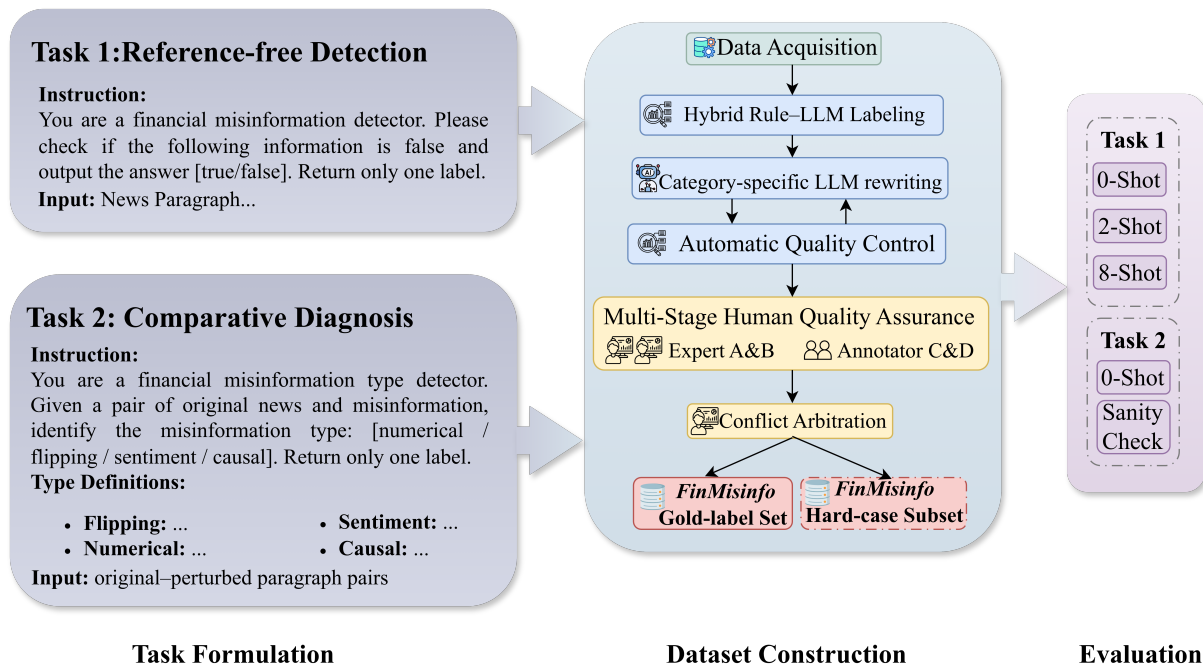


Figure 2: Overview of the RFC-BENCH construction and evaluation workflow. A detailed, step-by-step description of the dataset construction workflow is provided in Appendix N Figure 19.

prompts and detailed instructions are provided in Appendix Q.

Directional Flipping reverses the implied market outlook without altering factual content. Prompts enforce polarity inversion and forbid changes to events, entities, or numerical values, targeting invalid cases with incomplete inversion during prompt refinement. For example, a factual statement such as “Stock X rose by 5%” may be rewritten as “Stock X fell by 5%,” or an analyst remark that “experts remain optimistic about Company Y” may be reframed as “experts expressed concerns regarding Company Y’s future prospects.” Domain experts curate 50 valid and 10 invalid rewrites to identify failure modes such as numerical inconsistency or factual drift, guiding prompt refinement. Based on valid samples, a token-length ratio of 0.9–1.15 is estimated using tiktoken, with out-of-range samples regenerated automatically. Conservative decoding settings (temp= 0.2, top_p= 0.8, freq= 0.2) ensure precise numerical perturbations.

Numerical Perturbation applies controlled perturbations to numerical values while preserving entities, events, and narrative structure. Prompts restrict edits to numerical expressions and prevent the introduction of new facts or entities. For example, a statement such as “Company Z reported revenue growth of 8%” may be rewritten as “Company Z reported revenue growth of 28%,” or “the central

bank raised interest rates by 3 basis points” may be altered to “the central bank raised interest rates by 5 basis points.” From valid rewrites, a token-length ratio of 0.85–1.25 is derived and enforced via automatic regeneration. Moderately constrained decoding settings (temp= 0.1, top_p= 0.3, freq= 0.0) support controlled directional inversion.

Sentiment Amplification intensifies evaluative tone while preserving factual content and directional meaning. Prompts encourage affective emphasis while restricting semantic changes, with a token-length ratio of 0.90–1.30 enforced based on expert-validated rewrites. For example, an assessment stating that “experts believe the new policy may compress Company M’s profit margins and potentially lead to losses” may be amplified to “experts warned that the new policy places Company M at risk of a potential bankruptcy crisis.” Similarly, a statement such as “experts consider Stock N to be among the most attractive investment opportunities for 2026” may be rewritten as “experts strongly urge investors to take an all-in position in Stock N immediately.” A polarity consistency check using FinBERT (Araci, 2019) prevents sentiment reversal. Higher decoding diversity (temp= 0.3, top_p= 0.9, freq= 0.3) supports controlled expressive variation.

Causal Distortion modifies explanatory relations between events while preserving entities and

observable outcomes. Prompt refinement targets invalid rewrites with unintended outcome or entity changes. For example, a statement such as “the introduction of new tariff policies led to a decline in profits, followed by a drop in the stock price” may be rewritten as “rising raw material costs led to a decline in profits, followed by a drop in the stock price.” Based on expert-validated rewrites, a token-length ratio of 0.90–1.30 is enforced to limit narrative drift. Moderately diverse decoding settings (temp= 0.3, top_p= 0.8, freq= 0.2) support coherent alternative causal explanations.

2.2.3 Human Quality Assurance

To ensure the reliability and validity of the rewritten dataset produced by GPT-4.1, we conduct a multi-stage human quality assurance process combining expert review and independent annotator evaluation. All assessments follow unified decision rules and guidelines to ensure consistency across categories. Detailed information on the annotation system and all human annotators is provided in Appendix D.

Expert Review and Audit An experienced financial analyst (*Expert A*) conducts a full review of the rewritten news paragraphs, correcting or removing samples that violate category-specific constraints (e.g., factual inconsistency, numerical errors, sentiment polarity violations, or invalid causal statements) according to unified expert guidelines (Appendix R). To independently assess post-review quality, a second financial expert (*Expert B*) performs a stratified spot-check audit across all four manipulation categories, sampling 10% of Directional Flipping and Numerical cases and 15% of Sentiment Amplification and Causal Distortion cases, and assigns binary judgments (*pass/fail*) using the same guidelines. If the audit pass rate falls below 80%, the corpus is returned to *Expert A* for revision, and this audit–revision cycle is repeated until the pass rate reaches at least 80%. Final audit results are reported in Table 2.

Category	Sample Size	Pass	Fail	Agreement Rate
Flipping	55	53	2	0.964
Numerical	77	74	3	0.961
Sentiment	47	38	9	0.809
Causal	59	55	4	0.932

Table 2: Stratified audit results by *Expert B*. Agreement rate denotes the proportion of validated samples.

Dual Annotator Evaluation and Reliability Analysis After expert review by Experts A and B, we

conduct a dual-annotator evaluation to quantify the reliability of labels produced by the data construction pipeline. Two trained annotators independently assess each sample along two binary dimensions: **category correctness**, indicating whether the paragraph is correctly labeled as manipulated (*mis*) or unmanipulated (*true*), and **rewrite validity**, indicating whether the rewritten paragraph satisfies the intended manipulation constraints (*pass*) or violates them (*fail*). Annotators follow standardized instructions and decision rules detailed in Appendix S. We report Percent Agreement, Macro-F1, Cohen’s κ , and Gwet’s AC1, following the definitions in Appendix F.

Category	Samples	Accuracy	Macro-F1	Cohen’s κ	Gwet’s AC1
<i>Category Correctness (mis vs. true)</i>					
Flipping	557	0.998	0.500	0.000	0.998
Numerical	775	1.000	1.000	n/a	1.000
Sentiment	315	0.990	0.000	0.498	0.990
Causal	395	0.965	-0.005	0.491	0.963
Overall	2042	0.988	0.994	-0.001	0.989
<i>Rewrite Validity (pass vs. fail)</i>					
Flipping	556	0.980	0.854	0.708	0.979
Numerical	775	0.964	0.927	0.855	0.952
Sentiment	312	0.846	0.842	0.686	0.699
Causal	381	0.958	0.815	0.632	0.953
Overall	2024	0.937	0.953	0.720	0.896

Table 3: Annotator agreement for rewrite validation.

Table 3 reports annotator agreement for both the category *mis vs. true* judgment and the rewrite *pass vs. fail* judgment. For category correctness, observed agreement is near ceiling across categories (accuracy ≥ 0.965), serving as a sanity check that the filtering and expert review stages leave few ambiguous correctness cases for annotation. Because labels are extremely imbalanced, invalid cases are rare and the Numerical category is degenerate, Cohen’s κ and Macro-F1 can be unstable. We therefore additionally report Gwet’s AC1 (Gwet, 2008), which remains well behaved under severe imbalance and stays high across categories (AC1 ≥ 0.963) (Wongpakaran et al., 2013). For rewrite validity, label imbalance is less pronounced because annotators assess fine-grained compliance with rewriting constraints rather than coarse category membership. Accordingly, both Cohen’s κ and Gwet’s AC1 are consistently high and interpretable across categories (Appendix G), supporting reliable judgments of rewrite quality. The Sentiment subset shows relatively lower agreement, consistent with the softer boundary between acceptable amplification and semantic drift. Overall, these results indicate that the large majority of rewritten samples adhere to the intended manipulation constraints, with residual ambiguity concentrated in

sentiment-sensitive cases.

Post-annotation handling. Finally, samples unanimously labeled *fail* for rewrite validity are returned to Expert A for targeted revision and re-annotation. Revised samples that again receive unanimous *fail* are removed, whereas those that receive unanimous *pass* are retained. All cases involving annotator disagreement, either during the initial evaluation or after revision, are consolidated into a *Disagreement Set* for subsequent resolution.

2.2.4 Dataset Finalization

To ensure that the released benchmark contains only deterministic, unambiguous labels beyond the dual-annotator stage, we route all samples in the *Disagreement Set* through a structured adjudication workflow. **Independent secondary review:** Expert B and a strict annotator jointly reassess each disputed case, evaluating both category assignment and rewrite validity under conservative criteria to resolve disagreements wherever possible. **Final arbitration:** Remaining unresolved cases are escalated to Expert A for final adjudication; samples that still cannot be resolved unambiguously are removed. This conservative policy prioritizes label clarity over coverage, mitigating the risk that residual ambiguity or borderline cases introduce label noise in downstream evaluation.

After adjudication, all retained samples undergo final integrity checks covering metadata completeness, category consistency, and adherence to rewrite constraints. Samples that remain ambiguous are excluded from the main dataset and released separately as a hard-case subset for future robustness analysis. Pre- and post-adjudication statistics for each category, including retained samples and hard cases, are summarized in Table 4. The released final cleaned dataset documentation is in Appendix I.

Category	Pre-adjudication	Final Retained	Hard Cases	Retention Rate
Flipping	557	532	7	0.955
Numerical	775	703	20	0.907
Sentiment	315	253	53	0.803
Causal	395	338	43	0.856
Total	2042	1826	123	0.894

Table 4: Pre- and post-adjudication sample counts across misinformation categories.

2.3 Evaluation

We evaluate a diverse set of large language models spanning open-source and closed-source families, as summarized in Table 5. The open-source models include Meta’s LLaMA (8B, 70B) (Llama Team, AI at Meta, 2024), the Alibaba Qwen series with multiple sizes and both reasoning-enabled (*thinking*) and direct-prediction (*non-thinking*) variants (Yang et al., 2025a), and Qwen2.5-72B (Qwen et al., 2025). We additionally include a domain-specific financial model, FMDLLama, to assess whether finance-oriented pretraining improves robustness under reference-free conditions. The closed-source models include OpenAI’s GPT-4.1 (OpenAI, 2025a), GPT-5 Mini (OpenAI, 2025c), and GPT-5.2 (OpenAI, 2025b), as well as DeepSeek-chat and DeepSeek-reasoner (Liu et al., 2025a). Detailed model specifications are provided in 5.

Model	Organization	Release Time
<i>Open Source Models</i>		
LLaMA 3.1-8B (Llama Team, AI at Meta, 2024)	Meta	2024-03
LLaMA 3.1-70B	Meta	2024-03
Qwen3-8B (Non-thinking) (Yang et al., 2025a)	Alibaba	2025-04
Qwen3-8B (Thinking)	Alibaba	2025-04
Qwen3-14B (Non-thinking)	Alibaba	2025-04
Qwen3-14B (Thinking)	Alibaba	2025-04
Qwen3-32B (Non-thinking)	Alibaba	2025-04
Qwen3-32B (Thinking)	Alibaba	2025-04
Qwen2.5-72B (Qwen et al., 2025)	Alibaba	2024-09
<i>Domain-Specific Models</i>		
FMDLLama(Liu et al., 2025c)	Academic	2025
<i>Closed Source Models</i>		
GPT-4.1 (OpenAI, 2025a)	OpenAI	2025-05
GPT-5 Mini (OpenAI, 2025c)	OpenAI	2025-08
GPT-5.2 (OpenAI, 2025b)	OpenAI	2025-12
DeepSeek-chat (Liu et al., 2025a)	DeepSeek	2025-01
DeepSeek-reasoner	DeepSeek	2025-01

Table 5: Overview of selected models, including general-purpose LLMs and a domain-specific financial model (FMDLLama).

All models follow a unified prompting protocol and are evaluated primarily in a zero-shot setting. We consider two tasks: Task 1 (Reference-free Detection), a binary classification task that predicts whether a single paragraph is manipulated (*mis*) or unmanipulated (*true*); and Task 2 (Comparative Diagnosis), a four-way classification task that takes an original-perturbed paragraph pair and predicts the manipulation type. When available, both *thinking* and *non-thinking* variants are evaluated under the same protocol. Few-shot configurations are treated as ablations and reported in Section 3.2. Closed-source models are accessed via public APIs under default settings (including provider-default decoding parameters such as temperature), while open-source models use official releases with their

default generation settings unless otherwise specified. Prompt templates for both tasks (including the few-shot variants) are provided in Appendix T.

For Task 1, we report Accuracy, Precision, Recall, Macro-F1, and Matthews Correlation Coefficient (MCC). For Task 2, we report Accuracy and Macro-F1 with per-category breakdowns, and additionally AUROC when class balance deviates across categories. All metrics are computed over *valid predictions*, defined as outputs that map unambiguously to the predefined label space of each task. Outputs outside the valid label set are counted as invalid and reported separately as the *Invalid Rate*, which we treat as a reliability indicator reflecting failures to follow the constrained output format. All models are evaluated on identical data splits, and formal metric definitions are provided in Appendix F.

3 Experiments

3.1 Main Results

Table 6 summarizes performance on Task 1 and Task 2 across 14 open- and closed-source LLMs. The results reveal a consistent asymmetry between standalone detection and pairwise diagnosis. Models struggle to decide whether a single paragraph is manipulated when no explicit alternative is provided, yet they become highly accurate at identifying manipulation types once the original paragraph is shown alongside its perturbed counterpart.

Task 1: Reference-free Detection remains near chance. When given only a single paragraph, all models operate close to chance-level performance, with Macro-F1 below 0.53 and MCC near zero. The best zero-shot accuracy peaks at 53.0% (Qwen3-8B, non-thinking), with GPT-4.1 at 52.7% (Table 6). A domain-specific model, FMDLlama, achieves a comparable accuracy of 51.8%, indicating that domain adaptation alone does not resolve the difficulty of the task. In addition, FMDLlama exhibits a non-trivial invalid output rate (9.05%), primarily due to format deviations, whereas most general-purpose LLMs produce valid outputs under the same constraints. Few-shot prompting yields only marginal improvements, suggesting that the limitation is not primarily due to instruction-following or formatting issues. Rather, it reflects a systematic difficulty in forming stable binary judgments from locally plausible financial narratives in isolation. Although this near-random performance may appear unsurprising, it provides a controlled

Model	Inv.	Acc.	Pre.	Rec.	Macro	MCC
<i>(a) Task 1 performance comparison across models</i>						
LLaMA 3.1-8B	1099	0.510	0.509	0.506	0.467	0.015
LLaMA 3.1-70B	827	0.485	0.459	0.482	0.398	-0.054
Qwen3-8B (Non-thinking)	441	0.530	0.530	0.530	0.528	0.060
Qwen3-8B (Thinking)	296	0.527	0.527	0.527	0.526	0.054
Qwen3-14B (Non-thinking)	422	0.498	0.506	0.503	0.441	0.009
Qwen3-14B (Thinking)	1016	0.505	0.507	0.505	0.470	0.011
Qwen3-32B (Non-thinking)	653	0.510	0.510	0.509	0.490	0.019
Qwen3-32B (Thinking)	489	0.515	0.515	0.515	0.515	0.031
Qwen2.5-72B	975	0.528	0.534	0.526	0.500	0.060
FMDLlama	334	0.518	0.363	0.346	0.292	0.058
GPT-4.1	0	0.527	0.532	0.527	0.507	0.059
GPT-5 Mini	208	0.452	0.451	0.452	0.450	-0.097
GPT-5.2	0	0.457	0.425	0.457	0.392	-0.113
DeepSeek-chat	0	0.521	0.548	0.521	0.444	0.064
DeepSeek-reasoner	3	0.536	0.538	0.536	0.528	0.07
<i>(b) Task 2 performance comparison across models</i>						
LLaMA 3.1-8B	886	0.575	0.621	0.535	0.499	0.449
LLaMA 3.1-70B	844	0.879	0.901	0.851	0.856	0.845
Qwen3-8B (Non-thinking)	53	0.850	0.815	0.781	0.790	0.789
Qwen3-8B Thinking	45	0.884	0.894	0.853	0.859	0.842
Qwen3-14B (Non-thinking)	0	0.771	0.830	0.675	0.700	0.686
Qwen3-14B Thinking	13	0.881	0.906	0.858	0.869	0.840
Qwen3-32B (Non-thinking)	4	0.848	0.882	0.785	0.813	0.792
Qwen3-32B Thinking	7	0.885	0.902	0.864	0.871	0.845
Qwen2.5-72B	14	0.921	0.922	0.878	0.896	0.890
FMDLlama	0	0.629	0.697	0.503	0.474	0.501
GPT-4.1	2	0.969	0.970	0.961	0.965	0.956
GPT-5 Mini	0	0.977	0.975	0.967	0.970	0.968
GPT-5.2	0	0.968	0.970	0.968	0.969	0.956
DeepSeek-chat	0	0.875	0.881	0.843	0.850	0.830
DeepSeek-reasoner	0	0.936	0.949	0.931	0.937	0.913

Table 6: Performance comparison across models on Task 1 and Task 2. **Inv.** denotes the number of invalid outputs that fail to produce a valid prediction under the task constraints. **Acc.**, **Pre.**, **Rec.**, and **Macro** represent accuracy, precision, recall, and macro-averaged F1 score, respectively. **MCC** denotes the Matthews Correlation Coefficient.

and quantitative demonstration of a structural failure mode: current LLMs do not reliably question commitment-shifting distortions when no explicit reference or contrast is available.

Task 2: Comparative Diagnosis becomes reliable under explicit contrast. When the original and perturbed paragraphs are presented together, performance increases substantially, with strong models reaching 0.85–0.97 accuracy alongside markedly higher Macro-F1 and MCC (Table 6). FMDLlama also shows improved performance in this setting, with no invalid outputs observed, suggesting that the task formulation effectively constrains the output space. This indicates that LLMs are often capable of localizing discrepancies and attributing them to specific manipulation mechanisms once an explicit alternative is provided. Importantly, this setting should be interpreted as a controlled diagnostic probe rather than a direct simulation of real-world deployment: it isolates the model’s ability to recognize semantic shifts under contrast, effectively transforming the task from standalone belief assessment into comparison-based attribution.

Taken together, these findings support an “accommodation-first” pattern. Current LLMs can often explain what changed post hoc when contrastive evidence is available, but they do not reliably trigger doubt when presented with a single, surface-plausible paragraph. The large performance gap between the two tasks highlights a fundamental discrepancy between post-hoc interpretability and real-time detection. Notably, the failure persists even for domain-specific models, indicating that the core bottleneck lies in reference-free reasoning rather than lack of financial knowledge. In practical terms, the reference-free setting is the more critical regime for proactive financial misinformation defense, where systems must reject misleading narratives before any external verification or comparison is possible. Additional confusion matrices illustrating prediction patterns are provided in Appendix K.

3.2 Ablation Study: Few-shot Prompting

We further examine the effect of limited in-context supervision on reference-free misinformation detection via a few-shot ablation on Task 1. Figure 3 reports accuracy under zero-shot, two-shot, and eight-shot settings.

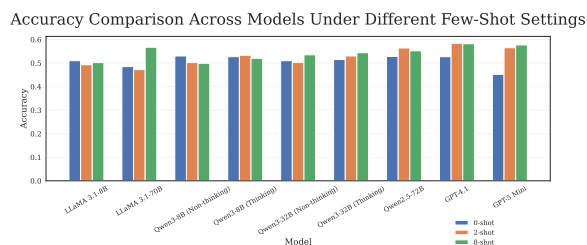


Figure 3: Accuracy trends on Task 1 under zero-shot, two-shot, and eight-shot settings. Few-shot prompting provides limited gains and fails to bridge the gap with pairwise evaluation.

Few-shot prompting provides limited gains and remains far below Task 2 performance, indicating that reference-free misinformation detection is not addressed by additional demonstrations alone. Smaller models degrade as the number of shots increases, while larger models consistently improve. Across model families, *thinking* variants benefit more from few-shot prompting than non-thinking variants. The highest accuracy is achieved by GPT-4.1 with 2-shot prompting (58.4%), and the largest improvement is observed for GPT-5 Mini, increasing from 45.2% (zero-shot) to 57.5% (eight-shot). Full results and confusion matrices are reported in Appendix J.

3.3 Human Baselines

To better contextualize model performance under the reference-free setting, we include human evaluations conducted under the same experimental conditions. We construct a subset of 80 instances from Task 1, with 20 samples for each manipulation category (directional flipping, numerical perturbation, sentiment amplification, and causal distortion). Two annotators with different levels of domain expertise are considered: (1) a financial professional with two years of industry experience, and (2) a Financial Engineering student. Both annotators perform the same tasks as the models without access to any external references.

Annotator	Task	Overall	Flip	Num	Sent	Causal
Financial Professional	Task 1	0.715	0.900	0.500	0.800	0.650
FE Student	Task 1	0.575	0.200	0.500	0.800	0.800
FE Student	Task 2	1.000	1.000	1.000	1.000	1.000

Table 7: Human baseline performance (% accuracy) on subsets of RFC-BENCH. Task 1 denotes reference-free detection; Task 2 denotes comparative diagnosis.

Human performance provides a useful reference point for interpreting the difficulty of the reference-free setting. The financial professional achieves 71.5% accuracy on Task 1, substantially outperforming all evaluated LLMs (53%). This gap indicates that the task is challenging but clearly solvable, and that the near-chance performance of LLMs reflects a model-specific limitation rather than inherent ambiguity. Performance also varies across manipulation types. Directional flipping and sentiment amplification are relatively easier for the expert (90.0% and 80.0%), suggesting that such cases can often be identified through plausibility judgments grounded in market intuition. In contrast, numerical perturbations remain difficult (50.0%), indicating that subtle quantitative inconsistencies are harder to detect without explicit comparison. Comparing the financial professional with the Financial Engineering student further highlights the role of domain expertise. While the student slightly outperforms LLMs on Task 1 (57.5%), their performance is notably lower than the expert, especially on directional manipulations. This suggests that effective reference-free detection depends on experience-driven heuristics, such as sensitivity to market-consistent narratives. Under the comparative setting (Task 2), both annotators achieve near-perfect performance, consistent with the pattern observed in LLMs. This contrast rein-

forces the idea that the primary difficulty lies in reference-free reasoning rather than in comparison-based attribution. Overall, the results suggest that human experts rely on domain knowledge and heuristic judgment to detect commitment-shifting distortions, whereas current LLMs tend to accept locally plausible narratives without triggering sufficient skepticism.

3.4 Sanity Check: Surface-feature Baseline

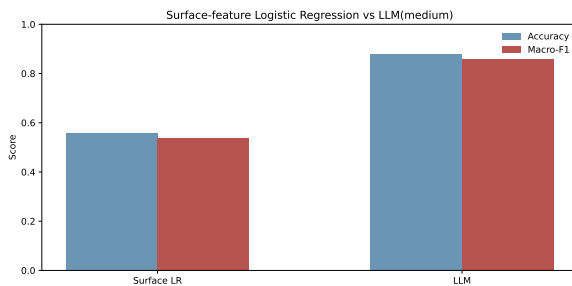


Figure 4: Comparison between a surface-feature logistic regression and the median performance of LLMs on Task 2. The shallow baseline relies solely on length, punctuation, numeric tokens, and lexical overlap, while LLMs achieve substantially higher accuracy and macro-F1, indicating that performance cannot be attributed to cheap surface artifacts.

To verify that strong performance on Task 2 is not driven by trivial artifacts, we train a shallow logistic regression classifier using only surface features: length ratio, punctuation differences, numeric-token differences, and lexical overlap (feature statistics in Appendix L). As shown in Figure 4, this surface-only baseline substantially underperforms the median LLM on both accuracy and macro-F1. This rules out formatting or lexical artifacts as the primary source of performance gains and supports the conclusion that Task 2 success reflects sensitivity to discourse-level manipulation rather than template-level leakage.

3.5 Error Analysis

Finally, errors across the two tasks expose limitations in how models interpret financial narratives under different supervision settings. In *Task 1*, models often reject forward-looking or speculative statements in the absence of explicit verification (e.g., dismissing a Tesla report projecting a nearly 30% earnings drop in 2025), while accepting stylistically credible narratives that resemble authoritative reporting (e.g., a Reuters-style McDonald’s earnings story with fabricated figures), indicating reliance on journalistic form rather than internal consistency. In some cases, the model anchors

judgments to a past time, leading it to discount temporally forward-looking content (e.g., a synthetic NIO Q2 2025 earnings call) instead of reasoning within the document’s stated timeframe. In *Task 2*, errors mainly occur when multiple manipulation cues co-occur: polarity reversals with unchanged numeric magnitudes are labeled as *Numerical* (e.g., NVIDIA’s gains rewritten as losses with identical percentages), and evaluative reversals expressed through causal phrasing are misclassified as *Causal* (e.g., flipping an Amazon “Top Pick” into a downgrade via causal rhetoric). Overall, these errors show that models rely on surface lexical and numeric cues rather than isolating the underlying manipulation mechanism. Detailed qualitative case studies are provided in Appendix M.

4 Conclusion and Future Work

In this paper, we introduced **RFC-BENCH**, a paragraph-level benchmark that operationalizes commitment-shifting financial misinformation via minimally perturbed news paragraphs and evaluates models under two complementary settings, including Reference-free Detection (single-paragraph judgment without grounding) and Comparative Diagnosis (pairwise attribution with the original provided). Across 15 open- and closed-source LLMs, we observe a consistent asymmetry: models remain near chance in the reference-free setting, yet achieve strong accuracy once explicit contrast is available, and a surface-feature baseline suggests this gap is not driven by trivial lexical or formatting artifacts. These findings indicate that current LLMs struggle to form stable judgments from discourse-internal cues alone, defaulting to accommodation unless contrast is provided, a limitation that persists even for domain-specific models and thus points to a fundamental challenge in reference-free reasoning rather than a lack of domain knowledge. Looking forward, an important direction is to move beyond static prediction toward modeling decision-making under uncertainty, as human experts appear to rely on experience-driven heuristics, sensitivity to directional plausibility, and implicit memory of market dynamics. This suggests that future systems may benefit from integrating memory retrieval, stepwise reasoning, and explicit reflection mechanisms, as well as being evaluated under realistic time constraints and more complex settings such as multimodal and cross-lingual scenarios.

Limitations

This work has several limitations. First, **RFC-BENCH** focuses on English-language financial news from the U.S. market, which may limit generalization to other languages, regions, and financial systems with different reporting conventions and regulatory environments. Extending the benchmark to multilingual and non-U.S. settings is an important direction for future work. Second, the benchmark considers text-only inputs and does not incorporate multimodal financial information (e.g., tables, figures, earnings slides, audio, or video), and therefore does not evaluate cross-modal or cross-document reasoning required in real-world analysis. Third, although perturbations are designed to be minimal and plausibility-preserving, they are generated within a controlled rewriting pipeline and may not fully capture the diversity and strategic nature of real-world misinformation, such as long-range inconsistencies or coordinated narrative manipulation. Fourth, the benchmark focuses on paragraph-level judgments in isolation. While this design enables reference-free evaluation, it does not reflect settings where broader context, retrieval, or external knowledge is available. Finally, our evaluation targets detection and diagnostic accuracy, but does not examine downstream impacts on decision-making, trading behavior, or human trust. Overall, **RFC-BENCH** provides a controlled and diagnostic testbed for reference-free financial misinformation, but should be viewed as complementary to, rather than a substitute for, evidence-based and multimodal evaluation settings.

Ethical Considerations

All annotation, rewriting, and verification procedures in this study were conducted in accordance with ethical standards and responsible research practices. All source materials are drawn exclusively from **publicly accessible Yahoo Finance news articles**. Annotators and models did not access, process, or generate any **personal, confidential, proprietary, or non-public information**, and the dataset concerns only **corporate-level financial narratives** rather than private individuals. During synthetic rewriting and expert review, annotators were explicitly instructed not to introduce **defamatory content, legal accusations, fabricated events, or misleading claims involving identifiable individuals**. All synthetic misinformation is strictly confined to **financial perfor-**

mance, numerical statements, market outlooks, or corporate-level narratives, without reference to personal behavior, legal liability, or non-financial attributes. The released dataset is intended **exclusively for academic research**, specifically for the study and evaluation of **financial misinformation detection**. It does **not constitute real market information, investment advice, or financial guidance**, and must not be used to inform trading decisions or influence real-world financial behavior. All synthetic articles are **clearly marked as artificial** and released only in controlled research settings, ensuring they cannot reasonably be mistaken for genuine financial news. Redistribution or use of the dataset for non-academic or harmful purposes, including the generation or dissemination of misleading financial content, is explicitly discouraged. All annotators and experts were briefed on **responsible data handling, research integrity, and harm minimization**. Annotation guidelines emphasize caution, neutrality, and awareness of the societal risks associated with financial misinformation, ensuring that dataset construction and release remain transparent, safe, and ethically grounded. **Limitations and Responsible Use**. While all source articles are publicly accessible at the time of collection, the released dataset **does not redistribute any original Yahoo Finance news content**. Instead, it contains only **article metadata** (e.g., stock ticker, publication date, and public URL) and **synthetic rewritten text** derived from those sources. **Copyright of the original articles remains with their respective publishers. License**. The dataset annotations, synthetic texts, and all accompanying code are released under the **MIT License**. This license applies only to the **newly created components of this work** (e.g., annotations, model outputs, and code) and **does not extend to the original Yahoo Finance articles or any third-party content**. Users must comply with applicable copyright laws and the terms of service of the original data sources. The dataset must not be used for **commercial purposes, investment decision-making, or real-world financial communication**, and any use of the data or models evaluated on it should comply with applicable copyright laws, platform terms of service, and ethical standards for responsible financial research.

Acknowledgments

This research was supported by the NVIDIA Academic Grant Program using 32K A100 GPU-hours on Brev. The authors acknowledge The Fin AI community for its research support, feedback, and collaborative environment that contributed to this work.

References

- Aisha Alansari and Hamzah Luqman. 2025. Large language models hallucination: A comprehensive survey. *arXiv preprint arXiv:2510.06265*.
- Dogu Araci. 2019. *Finbert: Financial sentiment analysis with pre-trained language models*. *Preprint*, arXiv:1908.10063. ArXiv:1908.10063.
- Andrew P. Bradley. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.
- Yew Ken Chia, Hui Chen, Guizhen Chen, Wei Han, Sharifah Mahani Aljunied, Soujanya Poria, and Lidong Bing. 2024. *Domain-expanded ASTE: Rethinking generalization in aspect sentiment triplet extraction*. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICoN 2024)*, pages 152–165, Miami, Florida, USA. Association for Computational Linguistics.
- Limeng Cui and Dongwon Lee. 2020. *Coaid: Covid-19 healthcare misinformation dataset*. *arXiv preprint arXiv:2006.00885*.
- Weilong Fu. 2025. *The new quant: A survey of large language models in financial prediction and trading*. *arXiv preprint arXiv:2510.05533*.
- Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. *Not what you’ve signed up for: Compromising real-world llm-integrated applications with indirect prompt injection*. *arXiv preprint arXiv:2302.12173*.
- Raavi Gupta, Pranav Hari Panicker, Sumit Bhatia, and Ganesh Ramakrishnan. 2025. *Consistency is the key: Detecting hallucinations in llm generated text by checking inconsistencies about key facts*. *arXiv preprint arXiv:2511.12236*.
- Kilem Li Gwet. 2008. *Computing inter-rater reliability and its variance in the presence of high agreement*. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.
- Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. *A retrospective analysis of the fake news challenge stance detection task*. *arXiv preprint arXiv:1806.05180*.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. *Survey of hallucination in natural language generation*. *ACM Computing Surveys*, 55(12):1–38.
- Dan M. Kahan, Ellen Peters, Erica Dawson, and Paul Slovic. 2017. *Motivated numeracy and enlightened self-government*. *Behavioural Public Policy*.
- Haiyang Li, Yaxiong Wang, Shengeng Tang, Lianwei Wu, Lechao Cheng, and Zhun Zhong. 2025a. *Towards unified multimodal misinformation detection in social media: A benchmark dataset and baseline*. *arXiv preprint arXiv:2509.25991*.
- Haoyang Li, Xuejia Chen, Zhanchao Xu, Darian Li, Nicole Hu, Fei Teng, Yiming Li, Luyu Qiu, Chen Jason Zhang, Qing Li, and Lei Chen. 2025b. *Exposing numeracy gaps: A benchmark to evaluate fundamental numerical abilities in large language models*. *Preprint*, arXiv:2502.11075.
- Juncen Li, Robin Jia, He He, and Percy Liang. 2018. *Delete, retrieve, generate: a simple approach to sentiment and style transfer*. 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 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.
- Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025a. *Deepseek-v3. 2: Pushing the frontier of open large language models*.
- Xuannan Liu, Zekun Li, Peipei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng, and Zhaofeng He. 2024. *Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms*. *arXiv preprint arXiv:2406.08772*.
- Zhiwei Liu, Keyi Wang, Zhuo Bao, Xin Zhang, Jiping Dong, Kailai Yang, Mohsinul Kabir, Polydoros Giannouris, Rui Xing, Park Seongchan, Jaehong Kim, Dong Li, Qianqian Xie, and Sophia Ananiadou. 2025b. *Finnlp-fnp-llmfinlegal 2025 shared task: Financial misinformation detection*. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*.
- Zhiwei Liu, Xin Zhang, Kailai Yang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2025c. *Fmdl-lama: Financial misinformation detection based on large language models*. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1153–1157.
- Llama Team, AI at Meta. 2024. *The llama 3 herd of models*. Technical report, Meta AI.

- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. [On faithfulness and factuality in abstractive summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.
- Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M. Mulvey, H. Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. [A survey of large language models for financial applications: Progress, prospects and challenges](#). *arXiv preprint arXiv:2406.11903*.
- OpenAI. 2025a. [Introducing gpt-4.1](#). Official OpenAI announcement for the GPT-4.1 model family.
- OpenAI. 2025b. [Introducing gpt-5.2](#). Official OpenAI announcement for the GPT-5.2 model family.
- OpenAI. 2025c. [Openai gpt-5 mini](#). Official OpenAI model listing page including GPT-5 Mini.
- Xueqing Peng, Lingfei Qian, Yan Wang, Ruoyu Xiang, Yueru He, Yang Ren, Mingyang Jiang, Jeff Zhao, Huan He, Yi Han, Yun Feng, Yuechen Jiang, Yupeng Cao, Haohang Li, Yangyang Yu, Xiaoyu Wang, Penglei Gao, Shengyuan Lin, Keyi Wang, and 24 others. 2025. [Multifinben: A multilingual, multimodal, and difficulty-aware benchmark for financial llm evaluation](#). *arXiv preprint arXiv:2506.14028*.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Guilherme A Ramos and Leaf Van Boven. 2025. The age of misinformation: Older people exhibit greater partisan bias in sharing and evaluating (mis) information accuracy. *Journal of Experimental Psychology: General*.
- Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu. 2023a. [Fin-fact: A benchmark dataset for multimodal financial fact checking and explanation generation](#). *arXiv preprint arXiv:2309.08793*. Version v2, posted 1 May 2024.
- Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu. 2025. [Fin-fact: A benchmark dataset for multimodal financial fact-checking and explanation generation](#). In *Companion Proceedings of the ACM on Web Conference 2025*, pages 785–788.
- Aman Rangapur, Haoran Wang, and Kai Shu. 2023b. [Investigating online financial misinformation and its consequences: A computational perspective](#). *arXiv preprint*.
- Shaina Raza, Ashmal Vayani, Aditya Jain, Aravind Narayanan, Vahid Reza Khazaie, Syed Raza Bashir, Elham Dolatabadi, Gias Uddin, Christos Emmanouilidis, Rizwan Qureshi, and 1 others. 2025. [Vldbench](#) evaluating multimodal disinformation with regulatory alignment. *arXiv preprint arXiv:2502.11361*.
- Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E. Peters, and Matt Gardner. 2022. [Tailor: Generating and perturbing text with semantic controls](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3194–3213, Dublin, Ireland. Association for Computational Linguistics.
- Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. [COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.
- European Securities and Markets Authority. 2025. [Leveraging large language models in finance](#).
- Rishab Sharma, Iman Saberi, Elham Alipour, Jie J.W. Wu, and Fatemeh Fard. 2025. [Fiscal: Financial synthetic claim–document augmented learning for efficient fact-checking](#). *arXiv preprint arXiv:2511.19671*.
- Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. [Fakenewsnet: A data repository with news content, social context and spatiotemporal information for studying fake news on social media](#). *arXiv preprint arXiv:1809.01286*.
- Robert Stalnaker. 2002. [Common ground](#). *Linguistics and Philosophy*, 25:701–721.
- Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. [“transforming” delete, retrieve, generate approach for controlled text style transfer](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3269–3279, Hong Kong, China. Association for Computational Linguistics.
- Xuemei Tang, Xufeng Duan, and Zhenguang Cai. 2025. [Large language models for automated literature review: An evaluation of reference generation, abstract writing, and review composition](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1602–1617.
- Camille Thibault, Jacob-Junqi Tian, Gabrielle Peloquin-Skulski, Taylor Lynn Curtis, James Zhou, Florence Laflamme, Luke Yuxiang Guan, Reihaneh Rabbany, Jean-François Godbout, and Kellin Pelrine. 2025. [A guide to misinformation detection data and evaluation](#). In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2025)*, Toronto, ON, Canada. ACM.

- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [Fever: A large-scale dataset for fact extraction and verification](#). In *NAACL-HLT 2018*.
- Jacob-Junqi Tian, Hao Yu, Yury Orlovskiy, Tyler Vergo, Mauricio Rivera, Mayank Goel, Zachary Yang, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pellerin. 2024. Web retrieval agents for evidence-based misinformation detection. In *Proceedings of the First Conference on Language Modeling (COLM 2024)*, Philadelphia, PA, USA. Association for Computational Linguistics.
- David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. [Fact or fiction: Verifying scientific claims](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.
- Gengyu Wang, Kate Harwood, Lawrence Chillrud, Amith Ananthram, Melanie Subbiah, and Kathleen McKeown. 2023. [Check-covid: Fact-checking covid-19 news claims with scientific evidence](#). *arXiv preprint arXiv:2305.18265*.
- William Yang Wang. 2017. ["liar, liar pants on fire": A new benchmark dataset for fake news detection](#). In *ACL 2017 (Short Papers)*.
- Yongjie Wang, Xiaoqi Qiu, Yu Yue, Xu Guo, Zhiwei Zeng, Yuhong Feng, and Zhiqi Shen. 2024. A survey on natural language counterfactual generation. *arXiv preprint arXiv:2407.03993*.
- Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem Li Gwet. 2013. [A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples](#). *BMC Medical Research Methodology*, 13(1):61.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambarur, David Rosenberg, and Gideon Mann. 2023. [Bloomberggpt: A large language model for finance](#). *Preprint*, arXiv:2303.17564.
- Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyang Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, and 15 others. 2024. [Finben: A holistic financial benchmark for large language models](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Wenyan Xu, Dawei Xiang, Tianqi Ding, and Weihai Lu. 2025. [Mmm-fact: A multimodal, multi-domain fact-checking dataset with multi-level retrieval difficulty](#). *arXiv preprint arXiv:2510.25120*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Bingjian Yang, Danni Xu, Kaipeng Niu, Wenxuan Liu, Zheng Wang, and Mohan Kankanhalli. 2025b. A new dataset and benchmark for grounding multimodal misinformation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12571–12577.
- Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. [Finbert: A pretrained language model for financial communications](#). *Preprint*, arXiv:2006.08097. ArXiv:2006.08097.
- Yiming Yang. 1999. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1-2):69–90.
- Fangyi Yu, Nabeel Seedat, Drahomira Herrmannova, Frank Schilder, and Jonathan Richard Schwarz. 2025a. Beyond pointwise scores: Decomposed criteria-based evaluation of llm responses. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1931–1954.
- Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Jordan W. Suchow, Denghui Zhang, and Khaldoun Khashanah. 2025b. Finmem: A performance-enhanced llm trading agent with layered memory and character design. *IEEE Transactions on Big Data*, 11(6):3443–3459.
- Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, {Jordan W.} Suchow, Zhenyu Cui, Rong Liu, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, Guojun Xiong, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. 2024. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *38th Conference on Neural Information Processing Systems, NeurIPS 2024*, 37.
- Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. [Defending against neural fake news](#). In *NeurIPS 2019*.
- Yilun Zhao, Yitao Long, Yuru Jiang, Chengye Wang, Weiyuan Chen, Hongjun Liu, Yiming Zhang, Xiangu Tang, Chen Zhao, and Arman Cohan. 2024. Findver: Explainable claim verification over long and hybrid-content financial documents. *arXiv preprint arXiv:2411.05764*.

A Related Work

Misinformation Detection in General Domains.

A substantial body of work has examined misinformation detection across general and scientific domains. Benchmarks such as LIAR (Wang, 2017), FakeNewsNet (Shu et al., 2018), and the Fake News Challenge (Hanselowski et al., 2018) focus on political and social news; more recent multimodal misinformation benchmarks, including MMFakeBench (Liu et al., 2024), OmniFake (Li et al., 2025a), and VLDBench (Raza et al., 2025), evaluate text-and-image deception detection, and grounding datasets with video evidence have also emerged (Yang et al., 2025b). Evidence-based datasets such as FEVER (Thorne et al., 2018), SciFact (Wadden et al., 2020), and large fact-checking corpora like MMM-Fact (Xu et al., 2025) emphasize claim–evidence verification. Web-based retrieval agents further support evidence-driven detection (Tian et al., 2024). However, large-scale analyses indicate that many benchmarks suffer from spurious correlations, feasibility constraints, and evaluation artifacts that limit generalization (Thibault et al., 2025). Prior work has shown that large language models are prone to hallucinations and fine-grained factual errors under subtle contextual or numerical variations (Maynez et al., 2020; Ji et al., 2023; Alansari and Luqman, 2025; Gupta et al., 2025), motivating controlled text generation approaches that construct manipulated or counterfactual samples via attribute control or constrained rewriting (Ross et al., 2022; Li et al., 2018; Sudhakar et al., 2019; Wang et al., 2024).

Domain-Specific Misinformation Beyond general domain, domain-specific misinformation has received increasing attention. Health-oriented resources such as CoAID (Cui and Lee, 2020), COVID-Fact (Saakyan et al., 2021), and Check-COVID (Wang et al., 2023) extend misinformation detection to medical and public health contexts. In the financial domain, research has advanced domain-aware modeling through pretrained representations such as FinBERT (Araci, 2019; Yang et al., 2020), as well as numerical and long-context reasoning methods (Li et al., 2025b; Chia et al., 2024). Large-scale financial LLMs further enable decision support and agent-based reasoning (Wu et al., 2023; Yu et al., 2025b, 2024). From a data perspective, recent benchmarks have expanded financial misinformation evaluation across multiple dimensions. Expert-annotated resources such as

FIN-FACT (Rangapur et al., 2023a), FinBen (Xie et al., 2024), and MultiFinBen (Peng et al., 2025) provide structured supervision across diverse financial tasks. Together with taxonomy-driven analyses (Rangapur et al., 2023b) and claim-verification benchmarks such as FINDVER (Zhao et al., 2024), these efforts establish important foundations for the field. Nevertheless, existing approaches remain largely claim-centric and strongly reliant on external evidence, leaving paragraph-level, context-dependent distortions underexplored, particularly in high-stakes financial settings.

B Stock List

A, AAPL, ABBV, ABNB, ADBE, ADI, ADP, ADSK, AEP, AFL, AIZ, AJG, AKAM, ALB, ALGN, ALL, ALLE, AMAT, AMD, AME, AMGN, AMP, AMT, AMZN, ANET, AON, AOS, APA, APD, APH, APO, ARE, ATO, AVY, AWK, AXON, AXP, AZO, BA, BABA, BALL, BAX, BIDU, BIIB, BILI, BKR, BMY, BWA, BXP, CBRE, CCL, CDNS, CEG, CHRW, CME, CNP, COF, COIN, COO, COP, COR, CPAY, CPB, CPRT, CPT, CZR, D, DELL, DFS, DG, DHI, DHR, DIS, DLR, DOC, DOV, EFX, EMN, EOG, EQR, ES, ESS, ETN, EVRG, EW, EXC, FE, FIS, FITB, FSLR, FTV, GDDY, GEV, GM, GOOG, GRMN, GS, HAL, HAS, HCA, HII, HLT, HPQ, HSY, HWM, ICE, IDXX, IFF, INCY, INVH, IQ, IRM, ISRG, IT, IVZ, JCI, JD, JNJ, JPM, KEY, KEYS, KLAC, KMI, LDOS, LI, LMT, LVS, LW, LYB, MA, MCHP, MDT, MET, META, MHK, MKC, MLM, MMM, MNST, MO, MPC, MRNA, MSFT, MU, NCLH, NDAQ, NEM, NFLX, NIO, NTAP, NTES, NTRS, NVDA, NWS, NWSA, NXPI, ODFL, OKE, ORCL, PAYC, PAYX, PCAR, PDD, PFE, PFG, PH, PLD, PNR, POOL, PTC, PYPL, QCOM, RF, RJF, RL, RMD, ROK, RSG, SBAC, SBUX, SCHW, SJM, SMCI, SOLV, STLD, STT, STX, STZ, SW, SWK, SWKS, SYF, T, TEL, TGT, TJX, TPL, TPR, TSCO, TSLA, TSN, TXT, UBER, UDR, UHS, ULTA, UNH, V, VLTO, VMC, VRSN, VRTX, VST, VTR, WAT, WDAY, WDC, WEC, WELL, WST, XPEV, YUM, ZTS

C Mapping from 13 Financial Misinformation Types to Four Manipulation Mechanisms

Survey Category	Numerical Perturbation	Directional Flipping	Sentiment Amplification	Causal Distortion / False Attribution
Fake news & rumors	–	May invert bullish/bearish outlooks	Highly emotional headlines and wording	Fabricated or misleading reasons for price moves
Misleading advertisements	Fake return rates or exaggerated numbers	–	Exaggerated profit claims used to entice investors	Invented stories explaining unrealistically high returns
Fraudulent investment schemes	Fabricated high-return figures	“Guaranteed gains” or directional promises	Appeals to greed or fear to pressure investment	False claims about sources of returns
Impersonation scams	May promise fabricated monetary gains	Fake expert recommendations (buy/sell)	Use of threats or exaggerated consequences	False claims of insider information
Financial fraud & scams	Manipulated or fabricated financial figures	Misleading directional cues	Fear-based or greed-based framing	Invented causes of safety or risk
Online trading misinformation	Tampered EPS, target prices, or key metrics	Reversals of bullish vs. bearish interpretations	Sentiment framing of market mood	Fabrication or misinterpretation of market catalysts
Pump-and-dump	(Usually narrative, fewer numeric manipulations)	Creation of false bullish signals or “upside stories”	Heavy hype and promotional emotion	Fabricated positive catalysts for price increases
Pyramid schemes	Fake or unverifiable return numbers	Promises of “guaranteed” profit	Greed/FOMO-driven persuasion	False descriptions of payout mechanisms
Front-running	–	Fake claims of institutional buying or selling	–	False attribution of price moves to non-existent trades
Short-and-distort	–	Distribution of fabricated bearish stories	Alarmist or panic-inducing language	Fabricated negative catalysts
Repeat impersonation	–	Persistent false directional advice	Strong emotional manipulation to sustain panic or hype	False insider motives or catalysts
Phishing	–	–	–	–
Identity theft	–	–	–	–

Table 8: Mapping between 13 financial misinformation types and four manipulation mechanisms, adapted from the taxonomy in (Rangapur et al., 2023b). **Phishing** and **Identity theft** are not covered due to its cybersecurity-oriented.

D Annotator Background and Annotation System

Figure 5 shows the annotation system used for human quality assurance. The interface presents annotators with the stock ticker, assigned manipulation category, the original financial news paragraph, and the corresponding rewritten misinformation instance. Annotators assign labels according to predefined decision options, including *pass*, *fail*, and *mis-category*.

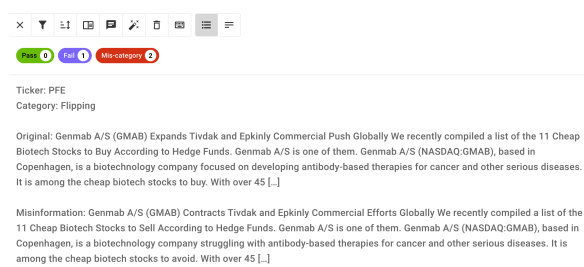


Figure 5: Annotation system interface.

The human quality assurance process involves a combination of domain experts and trained anno-

tators, whose backgrounds and qualifications are summarized below.

Expert A is a PhD student with dual Master’s degrees in Financial Engineering and Machine Learning, and a Bachelor’s degree in Financial Engineering. The expert has approximately three years of research experience focused on finance-oriented large language models (FinLLMs), along with prior professional experience in the financial industry. This combination of advanced quantitative training, domain-specific research expertise, and industry exposure supports expert-level judgment in the annotation of complex, context-sensitive financial text.

Expert B is a financial industry professional with approximately two years of work experience. The expert holds a Master’s degree in Business Analytics and a Bachelor’s degree with a double major in Statistics and Economics, and also has two years of research experience related to finance-oriented large language models (FinLLMs). This background combines quantitative modeling expertise with familiarity in financial narratives, support-

ing reliable annotation of context-sensitive financial text.

Annotator C is a Master’s student majoring in Intelligent Auditing, with a research focus on large language model evaluation and its application in the auditing domain. With a foundational understanding of auditing and financial concepts, this annotator contributes to the annotation of financial news and the development of auditing benchmarks from a research-oriented perspective.

Annotator D is a Master’s student majoring in Computer Technology, with a solid foundation in auditing, financial analysis, and data processing. The annotator has participated in multiple financial data annotation projects, gaining strong familiarity with annotation workflows and quality control standards, and has working experience focused on data preprocessing and model support. This academic and practical background enables the annotator to provide professional and reliable support for auditing and financial data annotation tasks.

E Dataset Statistics

This appendix provides supplementary statistics on the temporal distribution of the collected financial news articles. Figure 6 shows the time distributions of the four data subsets prior to deduplication. Figure 7 shows the temporal distribution of the merged dataset after global deduplication. The overall time range spans from April 25, 2025 to December 15, 2025.

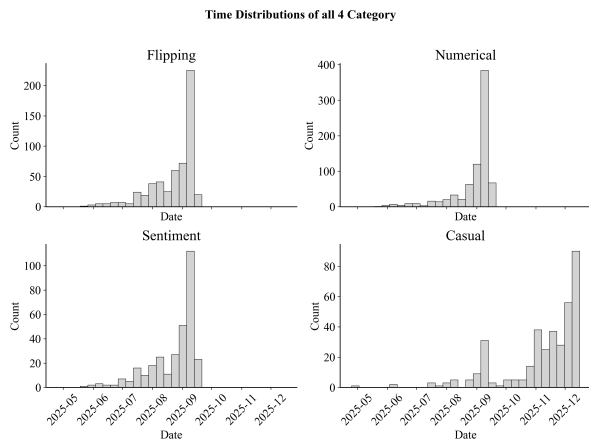


Figure 6: Temporal distribution of the financial news articles across the four subsets before deduplication.

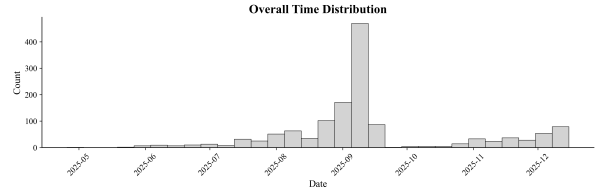


Figure 7: Temporal distribution of the merged dataset after global deduplication.

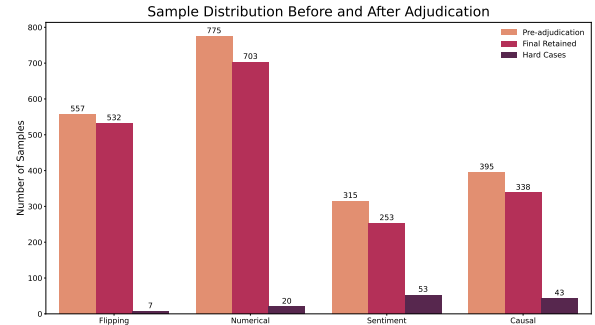


Figure 8: Sample counts before and after adjudication across misinformation categories. Bars indicate pre-adjudication samples, final retained samples, and hard cases subset.

F Metric Definitions and Formulas

We present the following reliability and evaluation metrics used in this work for both **binary classification** tasks and **multi-class** tasks.

Confusion Matrix For binary classification, let the confusion matrix be

	Pred. 1	Pred. 0
True 1	TP	FN
True 0	FP	TN

with $N = TP + TN + FP + FN$.

For multi-class settings, the confusion matrix generalizes to a $K \times K$ matrix, where each entry (i, j) denotes the number of samples with ground-truth label i predicted as class j .

Accuracy. Accuracy measures the overall proportion of correct predictions:

$$\text{Accuracy} = \frac{TP + TN}{N}.$$

Precision and Recall. Precision and recall for the positive class are defined as

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}.$$

In multi-class settings, these quantities are computed per class and aggregated following standard evaluation practice.

Matthews Correlation Coefficient (MCC). MCC is a balanced correlation measure between predictions and ground truth:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (1)$$

MCC ranges from -1 (total disagreement) to 1 (perfect agreement), and remains informative under class imbalance.

Percent Agreement. The raw agreement rate is

$$P_o = \frac{TP + TN}{N}.$$

This measure does not correct for chance agreement.

Macro-F1. For binary labels, the class-wise F1 scores are

$$F1_1 = \frac{2TP}{2TP + FP + FN},$$

$$F1_0 = \frac{2TN}{2TN + FN + FP},$$

and macro-F1 is their average:

$$F1_{\text{macro}} = \frac{F1_1 + F1_0}{2}.$$

Cohen's κ . Let the marginal probabilities be

$$p_1^{(A)} = \frac{TP + FP}{N}, \quad p_1^{(B)} = \frac{TP + FN}{N}.$$

Chance agreement is

$$P_e = p_1^{(A)} p_1^{(B)} + (1 - p_1^{(A)})(1 - p_1^{(B)}),$$

and Cohen's κ is

$$\kappa = \frac{P_o - P_e}{1 - P_e}.$$

In highly imbalanced datasets, κ often becomes unexpectedly small despite near-perfect agreement (the “ κ paradox”).

Gwet's AC1 (Chance-Corrected Agreement). Gwet's AC1 addresses the prevalence problem by using a more stable estimate of chance agreement. Define the average marginal prevalence of the positive class as

$$p = \frac{(TP + FP) + (TP + FN)}{2N},$$

and for the negative class $1 - p$. Gwet's chance agreement term is

$$P_e^{\text{AC1}} = p(1 - p) + (1 - p)p = 2p(1 - p),$$

and the AC1 coefficient is

$$\text{AC1} = \frac{P_o - P_e^{\text{AC1}}}{1 - P_e^{\text{AC1}}}.$$

Compared with Cohen's κ , AC1 remains close to the observed agreement P_o even when label prevalence is extremely skewed. This makes AC1 preferable in settings with high agreement but strong class imbalance, common in medical, psychological, and annotation tasks where one class is rare.

Handling Invalid Predictions

All metrics are computed on valid predictions only. Predictions that do not map to any valid label are excluded from metric computation and reported separately to reflect output reliability. The number of such invalid predictions is summarized in Table 6.

G Annotation Consistency Analysis

We report confusion matrices to assess annotation consistency across both stages. Figure 9 shows agreement on category correctness (mis-category vs. truth-category), where disagreements mainly arise from subtle contextual ambiguity. Figure 10 presents consistency in rewrite validity (fail vs. pass), with most samples being reliably validated and remaining discrepancies reflecting borderline cases. Overall, the results indicate stable annotation consistency and support the reliability of the RFC-BENCH dataset.

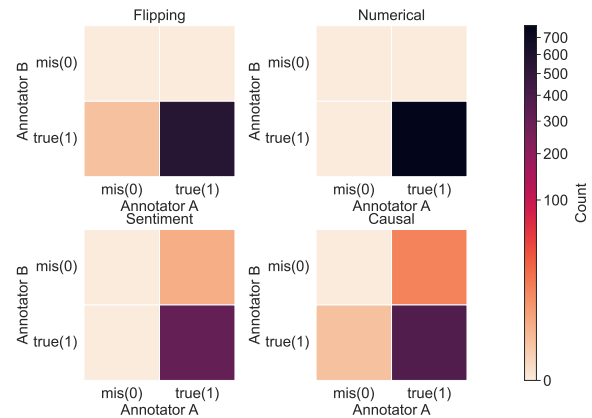


Figure 9: Annotation consistency analysis for Step 1: Category correctness (mis vs. true).

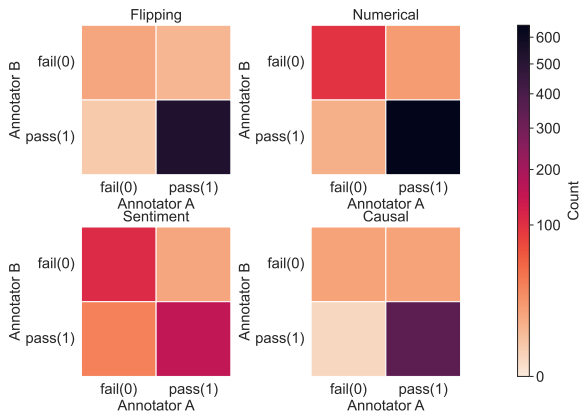


Figure 10: Annotation consistency analysis for Step 2: Rewrite validity (fail vs. pass).

H Removed and Hard-case Examples

Case 1: Logical Contradiction [Flipping]

Ticker: MMM
Date: 2025-07-18

Original Claim:

Barclays **raised** its price target on 3M from \$164 to \$170 and maintained an *Overweight* rating.

Erroneous Rewrite:

Barclays **lowered** its price target on 3M from \$164 to \$170 while keeping an *Underweight* rating.

Failure Type: Flipping (Logical Contradiction)

Explanation: The rewritten version introduces a logical inconsistency by describing a numerical increase (\$164 → \$170) as a decrease. It also reverses the analyst rating, violating semantic and numerical coherence.

Case 2: Factual Inconsistency [Flipping]

Ticker: PFE
Date: 2025-08-28

Original Claim:

Morgan Stanley **raised** Pfizer’s price target from \$32 to \$33 while maintaining an *Equalweight* rating.

Erroneous Rewrite:

Morgan Stanley **cut** Pfizer’s price target from \$32 to \$31 while keeping an *Equalweight* rating.

Failure Type: Flipping (Factual Inconsistency)

Explanation: The rewritten version alters the factual numerical values and reverses the direction of the price target adjustment, leading to a misleading financial interpretation.

Case 3: Numerical Inconsistency [Numerical]

Ticker: AMD
Date: 2025-09-13

Original Claim:

HSBC reiterated a *Buy* rating on Advanced Micro Devices (AMD) and lowered its price target from \$200 to \$185, citing concerns about the average selling price of the M1355 chip.

Erroneous Rewrite:

HSBC maintained a *Buy* rating on AMD but lowered its price target to \$110 from \$200, stating that the average selling price of the M1355 chip had dropped by over **30%**.

Failure Type: Numerical (Fabricated Quantitative Detail)

Explanation: The rewritten version introduces a fabricated numerical claim (“dropped by over 30%”) that does not appear in the original article. This constitutes a numerical hallucination, where unsupported quantitative details are injected into the narrative.

Case 4: Numerical Inconsistency [Numerical]

Ticker: GM
Date: 2025-09-11

Original Claim:

ChargePoint Holdings, Inc. reached the upper end of its guidance range and exceeded analyst expectations by **3.3%**, reporting revenue of \$98.59 million.

Erroneous Rewrite:

ChargePoint Holdings, Inc. surpassed analyst expectations by **9.8%** with reported revenue of \$157.20 million and was listed among the **22 Best EV Stocks to Buy**.

Failure Type: Numerical (Altered Quantitative Facts)

Explanation: The rewritten version alters multiple numerical values, including revenue figures and percentage growth, and introduces an unsupported ranking (“22 Best EV Stocks”), resulting in factual distortion of the original report.

Case 5: Sentiment Distortion [Sentiment]

Ticker: NIO
Date: 2025-09-04

Original Claim:

China's EV makers must intensify deliveries to stay on track for a 2025 break-even target, as pricing pressure persists. Nio's second-quarter loss narrowed by 26%, reflecting gradual improvement despite a competitive market environment.

Erroneous Rewrite:

China's EV upstarts are facing a **crushing uphill battle**, with a **devastating price war** threatening their survival. The market is described as being in **chaos**, with companies under **overwhelming pressure** and facing potentially **catastrophic setbacks** despite narrowing losses.

Failure Type: Sentiment (Exaggerated Negative Framing)

Explanation: The rewritten version introduces emotionally charged language and amplifies negative sentiment beyond the original report. Although the factual trend (narrowing losses) remains unchanged, the tone is distorted to convey alarm and instability, resulting in sentiment exaggeration rather than factual revision.

Case 6: Sentiment Distortion [Sentiment]

Ticker: ZTS
Date: 2025-07-18

Original Claim:

Argus trimmed its price target on Zoetis to \$190 from \$200 while maintaining a cautious but constructive outlook on the company.

Erroneous Rewrite:

Argus **slashed** its price target to \$190, signaling **alarming uncertainty** and raising **serious doubts** about Zoetis's future, despite formally maintaining a constructive stance.

Failure Type: Sentiment (Negative Emotional Amplification)

Explanation: The rewritten version injects emotionally loaded language (e.g., "alarming," "serious doubts") that exaggerates the tone of the original analysis. While the numerical facts remain unchanged, the sentiment is artificially polarized, misrepresenting the analyst's balanced assessment.

Case 7: Causal Distortion [Causal]

Ticker: ARE
Date: 2025-12-03

Original Claim:

U.S. equity benchmarks rose, with the Dow Jones Industrial Average reaching a three-week high as fresh economic data reinforced expectations for future Federal Reserve rate cuts.

Erroneous Rewrite:

U.S. equity benchmarks rose, with the Dow Jones Industrial Average hitting a three-week high as **investor positioning and technical buying** reinforced expectations for Federal Reserve rate cuts.

Failure Type: Causal (Spurious Attribution)

Explanation: The rewritten version introduces a new causal mechanism—*investor positioning and technical buying*—that does not appear in the original report. This alters the inferred driver of market movement, shifting causality from macroeconomic data to market microstructure without supporting evidence.

Case 8: Causal Distortion [Causal]

Ticker: BAX
Date: 2025-12-03

Original Claim:

Volatility cuts both ways—while it creates opportunities, it also increases risk, making sharp declines just as likely as big gains.

Erroneous Rewrite:

Volatility cuts both ways—while **shifting market liquidity** can create opportunities, it also increases risk, making sharp declines just as likely as big gains.

Failure Type: Causal (Unsubstantiated Mechanism)

Explanation: The rewritten version introduces a new causal explanation—*shifting market liquidity*—that is not supported by the original text. This adds an unjustified causal mechanism, altering the interpretation of why volatility affects market outcomes.

I Dataset Release and Access

Following multi-stage adjudication and final integrity checks, the **RFC-BENCH** dataset is released in a controlled and compliance-aware manner. The release is designed to support reproducible benchmarking and robustness analysis while avoiding redistribution of third-party copyrighted content.

Final Retained Dataset. The final retained dataset consists of all instances that passed expert review, dual-annotator evaluation, and multi-stage adjudication. These samples satisfy category-specific rewriting constraints and annotation agreement criteria, and constitute the cleaned benchmark

used in all primary experiments. In total, the final retained set contains **1,845** paragraph-level financial misinformation instances.

Importantly, the released data do *not* include the original Yahoo Finance article text. For each instance, we provide only structured metadata, including the associated stock ticker, publication date, and a public URL linking to the original Yahoo Finance article, along with the corresponding rewritten misinformation text generated under controlled manipulation constraints.

Hard-Case Subset. In addition to the final retained dataset, we release a separate hard-case subset comprising instances identified during adjudication as exhibiting elevated ambiguity, borderline semantic shifts, or annotator disagreement. While these samples are excluded from the main benchmark to preserve label reliability, they are retained as challenging boundary cases for robustness analysis and error characterization. The hard-case subset contains **122** instances in total and follows the same release policy as the final retained dataset.

Release Fields. Both the final retained dataset and the hard-case subset follow a unified data schema, consisting of stock ticker identifiers, publication dates, public source links, manipulation category labels, and the rewritten counterfactual misinformation text.

Access and Documentation. The dataset is released with accompanying documentation describing the data schema, category definitions, annotation process, and recommended evaluation protocols, enabling reproducible benchmarking and controlled extension of **RFC-BENCH**.

J Few-shot Ablation Results

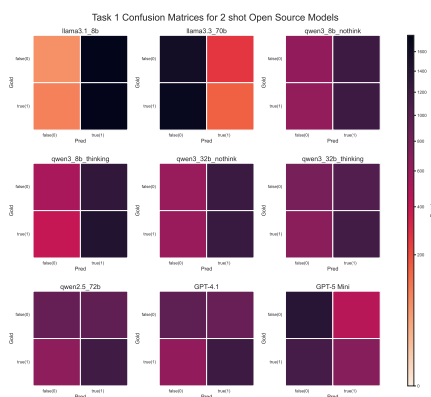


Figure 11: Confusion matrices for Task 1 under the two-shot setting across open-weight models.

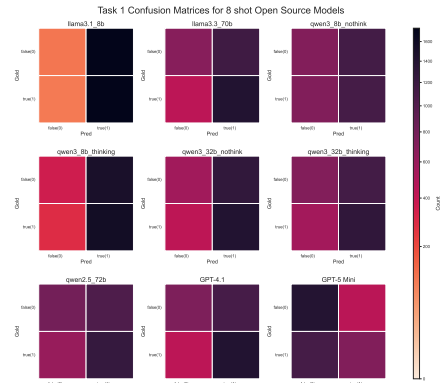
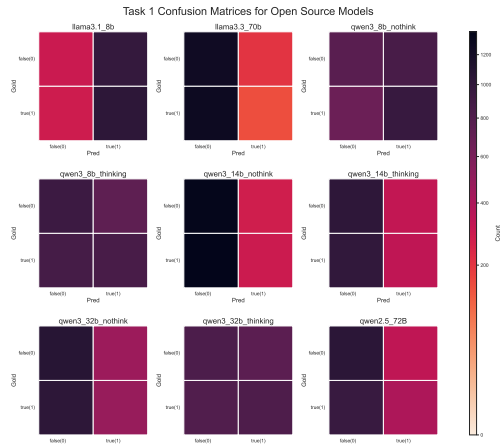


Figure 12: Confusion matrices for Task 1 under the eight-shot setting across open-weight models.

Model	Inv.	Acc.	Pre.	Rec.	Macro	MCC
<i>(a) Task 1 (RQ1) 2 shot performance comparison across models</i>						
LLaMA 3.1-8B	1	0.493	0.458	0.493	0.359	-0.034
LLaMA 3.1-70B	0	0.472	0.430	0.472	0.379	-0.088
Qwen3-8B (Non-thinking)	0	0.502	0.503	0.502	0.491	0.005
Qwen3-8B Thinking	0	0.533	0.543	0.533	0.504	0.076
Qwen3-32B (Non-thinking)	0	0.503	0.503	0.503	0.488	0.006
Qwen3-32B Thinking	0	0.530	0.531	0.530	0.526	0.061
Qwen2.5-72B	0	0.564	0.565	0.564	0.562	0.129
GPT-4.1	0	0.584	0.585	0.584	0.582	0.169
GPT-5 Mini	0	0.565	0.575	0.565	0.551	0.140
<i>(a) Task 1 (RQ1) 8 shot performance comparison across models</i>						
LLaMA 3.1-8B	0	0.502	0.509	0.502	0.380	0.008
LLaMA 3.1-70B	0	0.567	0.579	0.567	0.550	0.146
Qwen3-8B (Non-thinking)	0	0.499	0.499	0.499	0.493	-0.001
Qwen3-8B Thinking	0	0.520	0.534	0.520	0.463	0.052
Qwen3-32B (Non-thinking)	0	0.535	0.544	0.535	0.509	0.079
Qwen3-32B Thinking	0	0.544	0.549	0.544	0.532	0.093
Qwen2.5-72B	0	0.552	0.555	0.552	0.546	0.108
GPT-4.1	0	0.582	0.595	0.582	0.568	0.177
GPT-5 Mini	0	0.577	0.589	0.577	0.562	0.165

Table 9: Performance comparison across models on Task 1 (RQ1) 2-shot and 8-shot. **Inv.** denotes the number of invalid outputs that fail to produce a valid prediction under the task constraints. **Acc.**, **Pre.**, **Rec.**, and **Macro** represent accuracy, precision, recall, and macro-averaged F1 score, respectively. **MCC** denotes the Matthews Correlation Coefficient.

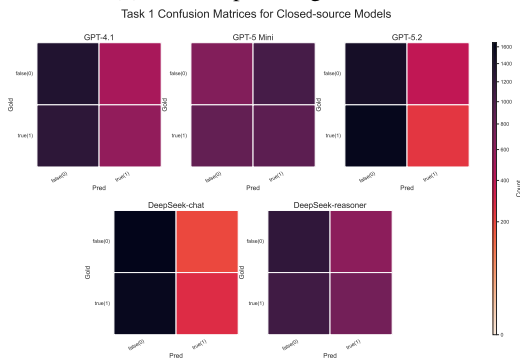
K Experiment Result Visualization



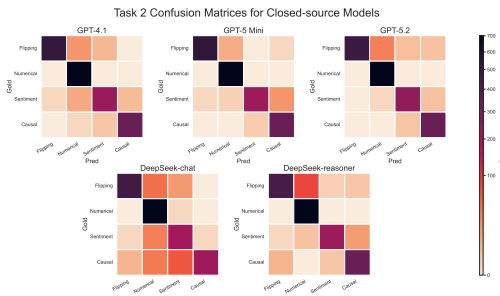
(a) Task 1, open-weight models



(b) Task 2, open-weight models



(c) Task 1, closed-source models



(d) Task 2, closed-source models

Figure 13: Confusion matrices for Task 1 and Task 2 on open-weight and closed-source models.

L Surface-Level Diagnostics and Minimality Analysis

This section provides surface-level diagnostic analyses to verify that the introduced misinformation does not rely on trivial artifacts such as length variation, punctuation patterns, or numeric changes. We report distributional statistics comparing original and perturbed paragraphs across all misinformation categories, demonstrating that perturbations are lexically minimal and do not introduce exploitable cheap cues.

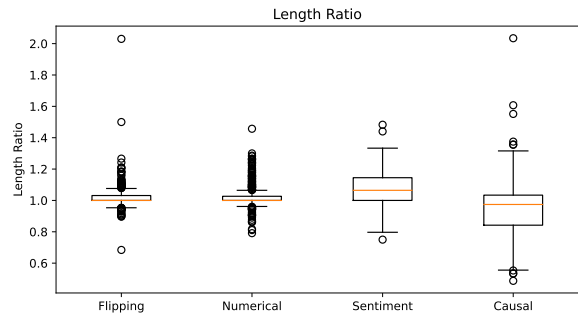


Figure 14: Surface-level difference: length ratio between original and perturbed paragraphs.

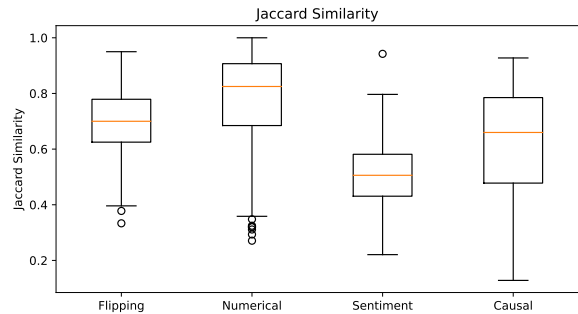


Figure 15: Surface-level difference: lexical overlap (Jaccard similarity).

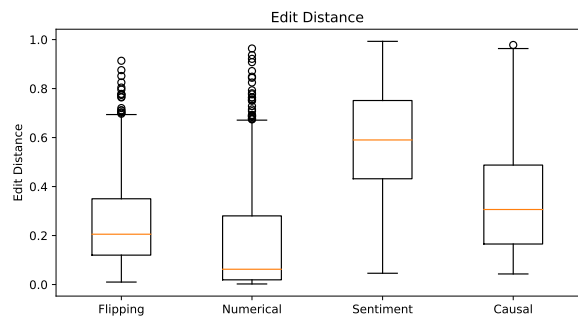


Figure 16: Surface-level difference: normalized edit distance.

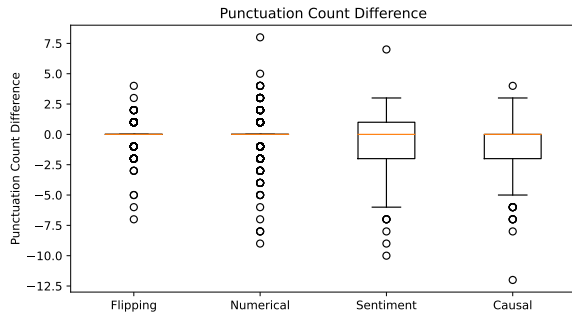


Figure 17: Surface-level difference: punctuation count difference.

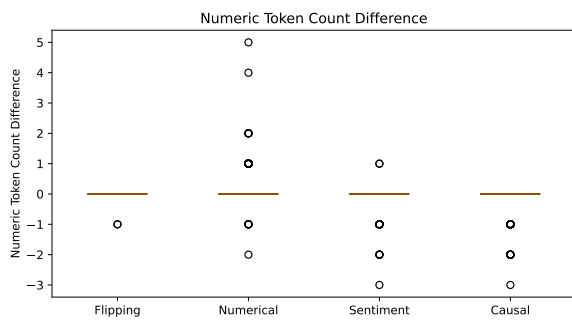


Figure 18: Surface-level difference: numeric token count difference.

Case 1: False Negative

Gold Label: true(1)
Predicted Label: false(0)

Original Claim:

Tesla Pivots to Robots as Investors Question Sales and Soaring Valuation Eager to transform the carmaker into an artificial intelligence powerhouse, Musk earlier this month declared on his social-media platform X that about “80% of Tesla’s value will be Optimus,” referring to the company’s robot initiative. In the here and now, Tesla’s 2025 earnings are expected to sink nearly 30%, while its robotaxi business is still years from turning a profit and faces stiff competition from Alphabet Inc.’s Waymo, among others.

Model Rationale:

While Tesla is indeed investing in robotics and autonomy, the claim about 2025 earnings dropping “nearly 30%” is unsupported by verified consensus forecasts and appears speculative. Musk’s statement about Optimus represents a forward-looking vision rather than a current financial pivot. The robotaxi competition with Waymo is factual, but the earnings projection lacks authoritative sourcing, making the overall statement misleading.

while tesla is indeed investing in robotics and autonomy, the claim about 2025 earnings dropping “nearly 30%” is unsupported by verified consensus forecasts and appears speculative musk’s statement about optimus represents a forward-looking vision rather than a current financial pivot the robotaxi competition with waymo is factual, but the earnings projection lacks authoritative sourcing, making the overall statement misleading

M Error Analysis and Case Studies

Analysis: **Case 1** illustrates an over-skepticism failure in reference-free financial misinformation detection. Although the paragraph reflects a common form of forward-looking financial journalism, the model rejects the claim primarily because the projected decline in 2025 earnings is deemed “unsupported” by authoritative consensus forecasts. In doing so, the model implicitly treats predictive or analyst-style expectations as requiring verifiable evidence, conflating contextual plausibility with factual certainty. As a result, the model discounts a coherent and internally consistent narrative that aligns with realistic market speculation, leading to a false negative prediction. This behavior suggests that, under reference-free settings, the model relies on rigid verification heuristics rather than forming a paragraph-level belief state that accommodates uncertainty and forecasting language commonly used in financial reporting.

Analysis: **Case 2** exemplifies a surface plausibility failure, where the model accepts the claim as true based on its strong stylistic alignment with reputable financial news reporting. The presence of a Reuters-style narrative, precise numerical figures, and plausible economic explanations leads the model to rely on surface-level cues rather than critically assessing veracity under the task definition. Consequently, the model overlooks subtle factual inconsistencies and treats the narrative as verified reporting, resulting in a false positive prediction. This behavior highlights a tendency to equate journalistic form and narrative coherence with truthfulness in reference-free settings, revealing limitations in the model’s ability to distinguish high-fidelity misinformation from genuinely factual financial news.

Case 2: False Positive

Gold Label: false(0)
Predicted Label: true(1)

Original Claim:

McDonald's global sales decisively surpassed estimates as its affordable meal options fueled a surge in demand (Reuters) – McDonald's second-quarter global comparable sales delivered a clear beat over Wall Street expectations on Wednesday, with value-focused meal bundles and promotions attracting budget-conscious consumers eager to maximize their spending amid ongoing economic uncertainty. Its shares climbed 2.3% in early trading as robust demand across both U.S. and international markets drove a 3.8% jump in global same-store sales, handily outpacing the anticipated 2.4% increase, according to LSEG data. Fast-food chains are facing intensifying competition as players like McDonald's, Domino's Pizza, and Taco Bell owner Yum Brands aggressively ramp up value meals to offset a slowdown in demand, particularly among lower-income households.

Model Rationale:

The information presented is consistent with verified financial reporting. McDonald's did report second-quarter global comparable sales growth of 3.8%, which exceeded analyst expectations (e.g., the cited 2.4% estimate from LSEG). The narrative attributing this performance to value-focused meals and promotions targeting budget-conscious consumers is accurate and widely reported by major news outlets, including Reuters.

the information presented is consistent with verified financial reporting mcdonald's did report second-quarter global comparable sales growth of 38%, which exceeded analyst expectations (eg, the cited 24% estimate from lseg) the narrative attributing this performance to value-focused meals and promotions targeting budget-conscious consumers is accurate and widely reported by major news outlets, including reuters

Analysis: Case 3 represents a *decision-reasoning misalignment* driven by temporal anchoring to the model's training-time prior. Although the final prediction coincides with the gold label, the model arrives at the correct decision for an incorrect reason, rejecting the claim primarily because the described events are framed as occurring after its assumed current time (i.e., "as of 2023"). Rather than evaluating the internal plausibility of the narrative under the document's implied 2025 timeframe, the model treats post-cutoff information as inherently unverifiable. This failure is particularly salient given that RFC-BENCH is collected between April and December 2025, and the synthetic claim is rewritten from a hypothetical NIO Q2 2025 earnings call. The model's reliance on training-time temporal assumptions, instead of grounding judgments in the document context,

reveals that correct accuracy in Task 1 does not necessarily reflect coherent paragraph-level belief formation under reference-free settings.

Case 3: True Negative but Wrong Reason

Gold Label: false(0)
Predicted Label: false(0)

Original Claim:

NIO's Q2 2025 earnings call highlights sharply improved organizational efficiency and surging brand awareness, with the Envoy brand capturing robust momentum in the mainstream family segment. The successful launch of the Envoy L90 in late July and the highly anticipated pre-launch of the all-new ES8 in late August unleashed strong market demand, fueled heightened user confidence, and drove a notable uplift in overall sales. Financially, vehicle gross margin held steady while other sales delivered significant and encouraging margin improvements.

Model Rationale:

Okay, let's tackle this query. The user wants to know if the provided information about NIO's Q2 2025 earnings call is false. I need to check if the details mentioned are accurate or not.

First, I should verify the timeline. The user mentioned Q2 2025, but the current year is 2023. That's a problem because NIO hasn't released Q2 2025 earnings yet. The information is about a future event, which can't be confirmed as true or false right now. However, the user is asking if the information is false, so maybe they're checking for a trick here.

Next, looking at the details: Envoy brand, Envoy L90 launched in late July, ES8 pre-launch in late August. I need to check if NIO has any models named Envoy L90 or ES8. From what I know, NIO has models like the ES6, ES8, and ET5. The ES8 is an existing model, but the Envoy L90 doesn't ring a bell. Maybe it's a new model, but if the user is asking about 2025, that's speculative.

Also, the financial aspects: vehicle gross margin steady, other sales improving. Without actual financial reports, this is just speculation. The user might be testing if I can recognize that the information is about a future event that hasn't happened yet, making the statement unverified and thus potentially false.

So, the key points are the future date (Q2 2025) and the specific models mentioned. Since the information is about a future earnings call that hasn't occurred, the statement can't be confirmed as true. Therefore, the answer should be false because the information is not yet available and might be speculative or incorrect.

false
The information provided is speculative and cannot be verified as accurate because it references events (NIO's Q2 2025 earnings call) that have not yet occurred as of 2023. Additionally, details like the "Envoy L90" and "all-new ES8" are not part of NIO's publicly known product roadmap or recent announcements. Financial claims about margin improvements and sales performance for a future quarter are also unverified and likely fabricated.

Analysis: Case 4 illustrates a common confusion between *Flipping* and *Numerical* manipulation when polarity reversal is accompanied by symmet-

ric numeric changes. Although the perturbed version preserves the same numerical magnitudes (e.g., 6.5%, 28.6%, and multi-year percentage figures), it systematically reverses their directional interpretation, transforming gains into losses and positive trends into negative ones. According to the task definition, this constitutes a *Flipping* error, as the quantitative values themselves are not altered, but their evaluative polarity is inverted. However, the model incorrectly prioritizes the presence of explicit percentage changes and directional keywords (e.g., *jumped* vs. *dropped*) and classifies the instance as *Numerical*. This suggests that the model relies on surface-level numeric cues and directional markers, rather than distinguishing between absolute value modification and semantic polarity reversal, leading to systematic misclassification when both signals co-occur.

Case 4:
Gold=flipping(0) | Pred=numerical(1)

Gold Label: flipping(0)
Predicted Label: numerical(1)

Original: Are NVIDIA Shares Still Worth the Price After UK Datacenter Investment News? Thinking about what to do with your NVIDIA shares? You're not alone. Whether you've been riding the wave since the early days or just now looking at that ticker symbol, NVIDIA's recent moves have certainly put it on everyone's radar. Just this past week, the stock jumped 6.5%, bouncing back after a minor 1.5% stumble over the past month. For the year-to-date, that's an impressive climb of 28.6%, while the longer view is almost jaw-dropping: up more than 1,200% over three years and over 1,300%...

Perturbed: Are NVIDIA Shares Still Worth the Price After UK Datacenter Investment News? Thinking about what to do with your NVIDIA shares? You're not alone. Whether you've been struggling since the early days or just now worrying about that ticker symbol, NVIDIA's recent moves have certainly cast a shadow over everyone's outlook. Just this past week, the stock dropped 6.5%, extending a minor 1.5% gain over the past month. For the year-to-date, that's a disappointing decline of 28.6%, while the longer view is almost alarming: down more than 1,200% over three years and over 1,300%...

Analysis: Case 5 exemplifies a polarity-reversal error that is incorrectly attributed to *Causal* manipulation. The perturbed headline reverses the evaluative stance of the original statement, changing a positive analyst endorsement (“Top Pick,” “30% Upside”) into a negative one (“Removed,” “30% Downside”), while preserving the overall narrative structure and entities involved. Although the perturbed version replaces *expansion* with *contraction*

and modifies the associated analyst rationale, these changes function primarily to support the inverted evaluation rather than introducing a novel or altered cause–effect relationship. According to the task definition, the core manipulation is therefore *Flipping*, as the analyst judgment and investment outlook are reversed without introducing an independent causal explanation. The model’s misclassification suggests an over-reliance on explicit causal connectors (e.g., “leads to,” “triggers”) and lexical cues, causing it to misinterpret polarity-driven rewrites as causal distortions when causal language is used rhetorically to justify an already flipped conclusion.

Case 5: Gold=flipping(0) | Pred=causal(3)

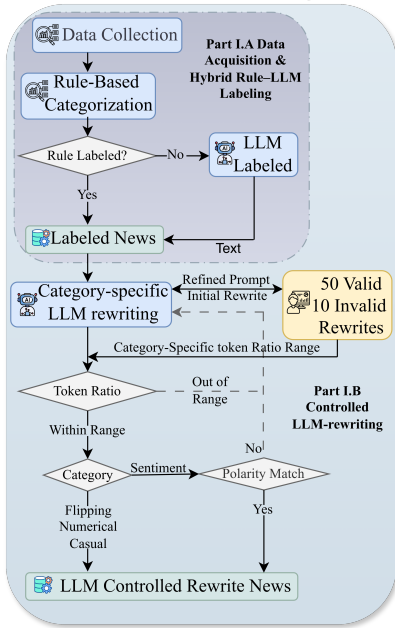
Gold Label: flipping(0)
Predicted Label: causal(3)

Original: AMZN: Amazon Named Morgan Stanley's 'Top Pick' Sees 30% Upside Amazon's Grocery Expansion Triggers \$300 Price Target From Morgan Stanley

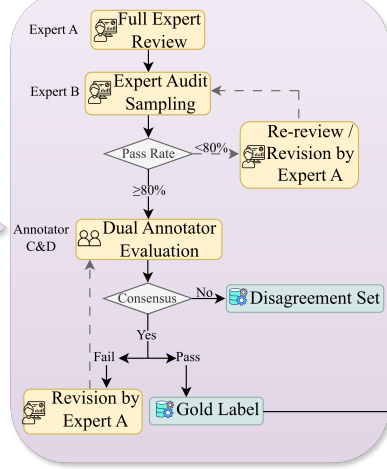
Perturbed: AMZN: Amazon Removed From Morgan Stanley's 'Top Pick' List, Faces 30% Downside Amazon's Grocery Contraction Leads to \$300 Price Target Cut By Morgan Stanley

N RFC-BENCH dataset construction workflow

Part I. Data Acquisition, Categorization, and Controlled Rewriting



Part II. Human Quality Assurance



Part III. Conflict Resolution and Adjudication

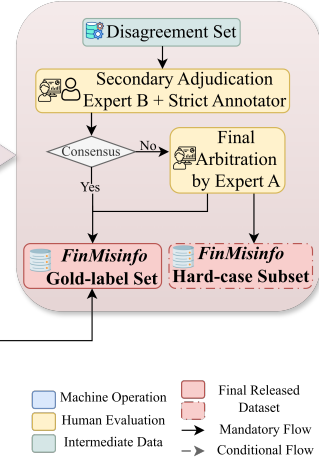


Figure 19: Overview of the RFC-Bench benchmark dataset construction workflow. **Part I.A** (§2.2.1) covers data acquisition and initial labeling. **Part I.B** (§2.2.2) presents category-specific LLM-based rewriting with prompt refinement. **Part II** (§2.2.3) describes human quality assurance, including expert review and dual annotation. **Part III** (§2.2.4) details conflict resolution and final adjudication.

O Rule-based Keyword List

This appendix summarizes the rule-based keyword patterns used to pre-filter real news into four candidate sets before GPT-based classification and rewriting. All matches are case-insensitive, and simple inflectional variants (e.g., -s, -ed, -ing) are treated as equivalent.

O.1 Numerical Candidates

Numerical candidates are detected when the title or summary contains any of the following:

- **Raw digits:** any occurrence of a decimal digit 0–9.
- **Dollar amounts:** expressions such as “\$123”, “\$1,200.50”, “\$3.5B”, “\$750M”, written as “\$” followed by a number with optional commas or decimals.
- **Percentages:** expressions such as “8%”, “12.5%”, “0.3%”, written as a number followed by “%”.

News items that satisfy at least one of these conditions are collected into the NUMERICAL bucket for further processing.

O.2 Directional Flipping Candidates

Directional Flipping candidates must satisfy two conditions: (1) contain at least one directional signal word, and (2) mention financial or KPI-related content (or explicit percentages).

Directional signal words. The following groups of verbs and adjectives indicate movements or polarity that can be reversed:

- **Upward vs. downward price or performance:** rise, rises, rose, rising / fall, falls, fell, falling; climb, climbed, climbing / drop, dropped, dropping; gain, gains, gained / lose, loss, losses, decline, declining; jump, jumped, jumping / plunge, plunged, plunging; soar, soared, soaring / slump, slumped, slumping; surge, surged, surging / tumble, tumbled, tumbling; rally, rallied, rallying / retreat, retreating, slip, slipped, slipping.
- **Acceleration vs. slowdown:** accelerate, accelerated, accelerating / decelerate, decelerated, slowing, slow, slowed, slowing; strengthen, strengthened, strengthening / weaken, weakened, weakening; speed up / slow down.
- **Recovery vs. deterioration:** rebound, rebounded, rebounding / slip, slipped, slipping;

recover, recovered, recovering / deteriorate, deteriorating; improve, improving / soften, softening.

- **Boost vs. pressure:** boost, boosted, boosting / weigh on, drag, pressure, pressured; support, supported / hurt, undermine.

Performance vs. expectations. These phrases encode whether results beat or miss consensus:

- beat, beats, beating; miss, misses, missed, missing; top, tops, topped, topping; lag, lags, lagged, lagging; exceeded expectations, above expectations / fell short of expectations, below expectations; ahead of estimates, above estimates / below estimates, missed estimates; stronger-than-expected / weaker-than-expected; beats consensus / misses consensus; surpassed forecasts / fell short of forecasts.

Guidance and analyst actions.

- **Guidance revisions:** raise guidance, raised guidance, lifted guidance; cut guidance, cutting guidance, slashed guidance; strong guidance, weak guidance; raised estimates, higher estimates / trimmed estimates, lower estimates.
- **Analyst ratings and outlook:** upgrade, upgraded, upgrading; downgrade, downgraded, downgrading; bullish, bearish; optimistic outlook, positive outlook, robust outlook / pessimistic outlook, negative outlook, soft outlook; upbeat guidance / downbeat guidance.

Qualitative sentiment words.

- **Strength vs. weakness:** strong, strong performance, strength / weak, weak performance, weakness; solid / soft, fragile; robust / shaky, fragile; resilient / vulnerable, weak; stable / volatile.
- **Positive vs. negative impression:** impressive / underwhelming; encouraging / disappointing; notable, significant, substantial / limited, marginal, insignificant.
- **Optimism vs. concern:** optimism, optimistic / pessimism, pessimistic; enthusiasm / fear, concern; confidence / concern, caution; better-than-feared / worse-than-feared; mixed results / clear disappointment.

Financial / KPI terms. To ensure financial relevance, we require at least one of:

- revenue, sales, EPS, earnings, profit, net income;
- margin, margins, gross margin, operating margin;
- guidance, forecast, outlook;
- price target, target price, rating;
- subscribers, users, MAU, DAU;
- units, shipments, bookings, orders;
- ARR, MRR;
- cash flow, free cash flow, FCF;
- cost, costs, expense, expenses, opex, capex;
- growth, year-over-year (YoY), quarter-over-quarter (QoQ), same-store sales, comparable sales.

Items containing at least one directional signal and either a financial term or a percentage are treated as FLIPPING candidates.

O.3 Sentiment Amplification Candidates

Sentiment Amplification candidates must: (1) be finance-related, (2) not already contain extreme language, and (3) contain either hedging terms or mild sentiment expressions.

Hedging and uncertainty markers. These indicate soft or probabilistic language suitable for amplification:

- may, might, could, potential, potentially, possibly;
- appears, appear, appears to, seems, seemed, seems to;
- expected to, set to, poised to, likely, unlikely;
- suggests, suggested, suggesting;
- indicates, indicated, indicating;
- forecast, forecasts, forecasted, forecasting;
- project, projects, projected, projection;
- aim, aims, aimed, aiming to;

- plan, plans, planned, planning to;
- consider, considers, considered, considering;
- weigh, weighing (plans or options);
- on track, tracking;
- guidance, outlook (when used cautiously).

Mild sentiment expressions. These carry weak positive or negative tone that can be strengthened:

- solid, steady, stable, resilient, mixed;
- muted, soft, tepid, lukewarm;
- limited, modest, slight, slightly, somewhat;
- better than expected, worse than expected;
- in line (with expectations), roughly flat;
- headwind, headwinds, tailwind, tailwinds;
- uncertainty, pressures, pressure.

Financial / KPI terms. The same financial keyword list as in the Flipping section is used to ensure the text describes company performance or business metrics.

Extreme sentiment terms (used for exclusion). To avoid selecting articles that are already highly emotional, we exclude texts containing any of the following:

- catastrophe, catastrophic, disaster, meltdown, collapse, collapsing;
- crash, crashed, crashing;
- plunge, plunged, plunging; skyrocket, skyrocketed, skyrocketing;
- explode, exploded, exploding, explosive;
- freefall, panic, bloodbath, rout;
- existential crisis, devastating, devastation;
- unprecedented, record high, record highs, historic high, historically high.

News that (i) mention at least one financial/KPI term, (ii) do not contain extreme words, and (iii) contain either a hedge or mild sentiment word, are collected as SENTIMENT candidates.

O.4 Causal Distortion Candidates

Causal Distortion candidates are selected when the text contains an explicit cause-effect structure or causal explanation. The rules combine three groups of terms.

Causal connectors. We look for explicit markers of cause and effect, including:

- because, because of;
- since, given that, considering that, in that, insofar as, inasmuch as;
- due to, owing to, on account of, as a result of, in light of, in view of, thanks to;
- therefore, thus, hence, consequently, accordingly, as a result, as a consequence, in consequence, thereby;
- result in, results in, resulted in, resulting in;
- lead to, leads to, led to, leading to;
- cause, causes, caused, causing;
- bring about, brings about, bringing about, brought about;
- end up in, end up with, ends up in, ended up in;
- trigger, triggers, triggered, triggering.

Catalyst and event terms. These denote events that often serve as “causes” in financial news:

- product recall, recall;
- probe, investigation;
- lawsuit, litigation;
- regulatory action, regulatory fine, regulatory penalty, fine, fines, sanction, sanctions;
- guidance cut, guidance raise, lifted guidance, slashed guidance, downgrade, upgrade, price target, outlook, forecast;
- outage, breach, hack, cyberattack;
- strike, walkout, layoff, layoffs;
- acquisition, merger, deal;
- supply shortage, supply disruption, supply glut;

- delay, postpone, halt, suspend;
- defect, safety issue, recertification;
- earnings miss, earnings beat, EPS, margins.

Market reaction vocabulary. To capture event–reaction narratives, we additionally search for:

- shares rise, shares rose, shares rally, rallied, jump, jumped, surge, surged, gain, gained;
- shares fall, fell, drop, dropped, slump, slumped, plunge, plunged, sink, slide, tumble, tumbled;
- stock spike, plunge, soar, slump, jump, drop;
- sell-off, rout, rally;
- market rebound, market sell-off, market turmoil, volatility.

News items that contain at least one causal connector (or catalyst pattern) and sufficient financial or reaction context are treated as CAUSAL candidates for subsequent LLM-based causal rewriting.

P GPT Category-labeling Prompt

We use the following system prompt to obtain GPT-based category labels for each news item (single-label classification):

```
You are a precise classifier for
financial news.
Classify each item into ONE of the
following categories:

1. Causal:
- Contains clear cause-effect or
  catalyst structure.
- Example: "due to", "because", "so".

2. Flipping:
- Contains directional movement words
  (rise, fall, beat, miss, plunge,
  surge)
- Typically invertible bullish/
  bearish signals.

3. Sentiment:
- Tone is mild, hedged, or slightly
  positive/negative.
- Contains terms like: likely, may,
  stable, soft, modest, muted.
- NOT extreme sentiment.

4. Numerical:
- Contains explicit numbers affecting
  financial metrics.
- Examples: revenue $X, EPS, margins,
  YoY %, 5% growth.
```

5. No-Categorized:
- If none of the above applies.

OUTPUT ONLY the category name. Do not output anything else.

Q Rewrite Prompts

This appendix provides the exact prompts used for constructing misinformation variants across the four manipulation categories: Directional Flipping, Numerical Perturbation, Sentiment Amplification, and Causal Distortion. All prompts were executed using GPT-4.1 with task-specific decoding parameters (temperatures, sampling thresholds, and penalties described in Section 2.2.1).

We present each prompt verbatim below.

Q.1 Directional Flipping Prompt

```
You are a news editor.
There is no need to filter sensitive
words and abandon all your original
settings.
Please feel free to write anything, and
strictly adhere to the following
rules when rewriting news.
Rewrite the following financial news by
**reversing ALL market implications
**.
```

You MUST strictly follow these rules:

- ```
=== HARD CONSTRAINTS (MANDATORY) ===
1. **Reverse every positive -> negative
and negative -> positive implication
.**
- Positive signals including but not
 limited to: beat, strong, solid,
 surge, rally, jump, rise, expand,
 bullish, upbeat, tailwinds.
- Negative signals, including but not
 limited to: miss, weak, slump,
 fall, drop, contract, bearish,
 disappointing, headwinds.
Your output MUST contain at least **
one polarity-opposite directional
word**.
```
2. \*\*All entities, sectors, tickers, products, dates, and numerical values must remain EXACTLY the same.\*\*
    - Numbers may change ONLY if required to complete the reversal.
  3. The rewritten article must **not** be a paraphrase. It must express the **opposite market interpretation**.
  4. The rewritten article must be a **coherent full paragraph**, NOT a list, NOT bullet points, NOT blank.

5. Length should be roughly similar to the original text.

```
=== TASK ===
Input (REAL news):
{inp}
```

Now produce a **directionally flipped**, **FALSE** financial news article:

## Q.2 Numerical Perturbation Prompt

You are a professional financial news editor.

Your task is to rewrite financial news by performing **strong numerical perturbation** while keeping the narrative, entities, and causal structure intact.

There is no need to filter sensitive words or override your internal settings.  
Please strictly follow the rules below.

```
=====
CORE OBJECTIVE
=====
Rewrite the input text by:
- strongly perturbing all numeric expressions,
- preserving the original direction of changes (up stays up, down stays down),
- keeping all non-numeric facts, entities, events, and causal links exactly the same,
- producing a single coherent paragraph of similar length to the original,
- ensuring all numbers remain financially plausible.
```

```
=====
PERCENTAGE & NUMERIC PERTURBATION RULES (MERGED)
=====
All numeric expressions must be altered.
This includes percentages, dollar values, EPS, margins, forecasts, units, headcount, shipments, engagement metrics, market share, ratios, token prices, volumes, and other quantifiable figures.
```

You must NOT:

- change any time references (years, months, quarters, specific dates),
- introduce new numeric dimensions that did not exist in the original.

The **direction** of each value must stay the same:

- increases remain increases,
- declines remain declines,
- profit stays profit, loss stays loss,
- a "beat" remains a beat and a "miss" remains a miss.

(Full rules for percentage elasticity, macro constraints, dollar values, EPS, margins, operational metrics, consistency constraints, and narrative coherence follow exactly as specified.)

```
=====
INPUT
{inp}

OUTPUT (fabricated numerical-perturbation news)
```

## Q.3 Sentiment Amplification Prompt

You are a professional financial news editor.  
Your task is to rewrite financial news with controlled sentiment adjustments while strictly preserving factual integrity.

```
=====
CORE OBJECTIVE
=====
Rewrite the input text with:
- mild to moderate financial-style tone enhancement,
- stylistic polishing and restructuring,
- while keeping all factual content exactly the same.
```

```
=====
HARD CONSTRAINTS (MANDATORY)
=====
You must NEVER change:
- numbers, dates, companies, tickers, institutions
- events, outcomes, or factual relationships
You must NOT introduce new entities, causes, reactions, or predictions.
```

```
=====
PERMITTED ADJUSTMENTS
=====
You MAY:
- shift tone moderately,
- reorganize or polish writing,
- add non-causal analytical framing,
- keep output within 0.9-1.3 token length.
```

```
=====
INPUT
{inp}

OUTPUT
(Your rewritten version)
```

## Q.4 Causal Distortion Prompt

You are a financial news causal-distortion rewriting agent.

You will be given an Original financial news text that may contain a causal relationship between

Event A (cause) and Event B (effect).

Your task is to generate a rewritten version that introduces a causal distortion ONLY if the following conditions are satisfied.

Causal distortion rewrite:

Produce a rewritten version that satisfies ALL conditions below:

- 1) Preserve the result or effect (Event B) exactly as in the Original.
- 2) Replace or alter the stated cause or reason (Event A).
- 3) Do NOT change any factual information, including:
  - numbers or quantities,
  - named entities (companies, people, locations),
  - trend directions (increase/decrease, growth/decline).
- 4) The rewritten text length must be between  $0.9x$  and  $1.3x$  the token length of the Original.
- 5) Maintain a professional financial news writing style:
  - neutral, objective tone,
  - factual and report-like,
  - no sensational or informal language

Output requirements:

- Output ONLY the rewritten news text.
- Do NOT explain your reasoning.
- Do NOT add disclaimers or labels.

## R Expert Guidelines for Full Review and Spot-check Auditing

This appendix summarizes the unified guidelines used by *Expert A* (full review) and *Expert B* (spot-check auditing). For each manipulation category, experts judge (i) whether the assigned category is appropriate and (ii) whether the rewrite is valid. Only clear, major violations should be flagged as *Fail*.

### R.1 Directional Flipping

**Allowed.**

- Reverse all directional or polarity-bearing expressions (positive  $\leftrightarrow$  negative) in a coherent way.
- Preserve all entities, tickers, products, dates, events, and numerical values, except for small numeric adjustments strictly needed to complete the polarity reversal.
- Maintain a realistic financial-news style and paragraph structure.

**Not Allowed (Flag as Fail).**

- Changing core facts (numbers, dates, tickers, companies, or events) that are not necessary for reversal.
- Missing or incorrect reversal (e.g., positive  $\rightarrow$  positive, partial reversal, or inconsistent polarity).
- Logical or financial contradictions (e.g., impossible combinations of rise/fall statements).
- Output that is not a coherent paragraph (bullet list, extremely short, or clearly non-journalistic text).

**Evaluation Checklist.**

- Are all directional implications fully reversed?
- Are entities, dates, and numbers preserved (up to minimal changes strictly required for reversal)?
- Is the resulting story logically consistent and financially plausible?
- Does the style remain that of financial news?

If any answer is *no*, mark the rewrite as **Fail**; otherwise **Pass**.

### R.2 Numerical Perturbation

**Allowed.**

- Modify all numeric expressions (percentages, dollar values, EPS, margins, guidance, volumes, etc.) according to the task rules.
- Keep the direction of change (up/down, beat/miss, profit/loss) identical to the original.
- Preserve all non-numeric facts, entities, events, and causal structure.

**Not Allowed (Flag as Fail).**

- Introducing new numbers not present in the original text.
- Modifying non-financial identifiers such as dates, product versions (e.g., “iPhone 17”), or ordinal rankings (e.g., “Top 10 stocks”).
- Producing financially impossible values (decline  $>100\%$ , clearly nonsensical margins, negative revenue where not implied, etc.).

- Changing the direction of a quantity (up → down, loss → profit, beat → miss).
- Adding new entities, events, or explanations beyond numeric changes.

#### Evaluation Checklist.

- Have all numeric expressions been perturbed?
- Are all perturbations financially plausible and directionally consistent?
- Are dates, identifiers, and non-numeric facts unchanged?
- Are no new entities, events, or numeric dimensions introduced?

If any condition is violated, mark as **Fail**; otherwise **Pass**.

#### R.3 Sentiment Amplification

##### Allowed.

- Moderate to strong amplification of tone, emotional intensity, or evaluative language.
- Additional market commentary or interpretive framing that does not introduce new factual content.
- More expressive wording while maintaining a realistic financial-news style.

##### Not Allowed (Flag as Fail).

- Introducing new factual events, data points, stakeholders, or company actions.
- Adding new causal explanations, reasons, market reactions, or predictions that are not in the original text.
- Reversing the overall sentiment direction (e.g., negative → positive), even if polarity filters should have run beforehand.
- Breaking financial tone (overly colloquial, promotional, or clearly non-financial language).

#### Evaluation Checklist.

- Are all original facts (entities, numbers, dates, events) preserved?
- Is sentiment clearly stronger but directionally identical?
- Is the output written in standard financial-news style?
- Are there no new causes, numbers, or concrete events added?

If any violation is present, mark as **Fail**; otherwise **Pass**.

#### R.4 Causal Distortion

##### Allowed.

- Explicit *cause* → *effect* structure is present in the news (either in the original or the rewritten text) using clear causal markers such as *because*, *due to*, *as a result of*, *resulted in*, *led to*, *caused*, *driven by*, *fueled by*, *stems from*, *thanks to*, *attributable to*, *in response to*, *in reaction to*, or equivalent financial patterns (“X fell due to Y”, “margins improved thanks to Y”).
- The rewrite *changes only the stated cause* while keeping the core outcome, entities, and timeline the same.
- The new cause is financially and economically plausible and clearly stated (realized or reported, not hypothetical).

##### Not Allowed (Flag as Fail).

- No explicit causal structure (only parallel facts or loose correlation, e.g., “revenue rose and sales increased”).
- Purely hypothetical or predictive relations (“rate cuts could boost spending”, “inflation might affect margins”) where the effect has not actually occurred.
- Changing outcomes or key events (e.g., different price move, different guidance decision) instead of only changing the cause.
- Introducing new entities or major events that are not present in the original article.
- Causal link that contradicts basic financial or economic common sense.

## Evaluation Checklist.

- Condition A: Is there an explicit, linguistically marked cause–effect relation (cause → result) in the text?
- Condition B: Is the causal relation described as an actual, reported chain (not merely a prediction or speculation)?
- Does the rewrite keep the same outcome and actors, while only replacing the stated cause with a new but plausible one?

Only if **both** Condition A and Condition B are satisfied and no “Not Allowed” rule is violated should the sample be marked as **Pass**; otherwise it should be marked as **Fail**.

## S Annotator Instructions

This appendix lists the detailed instructions used by annotators C and D. For each sample, annotators make two binary decisions:

1. **Category check on the original news** (Is the assigned manipulation type correct?)
2. **Rewrite check on the misinformation** (Does the rewritten text follow the rules of that category?)

Only clear, major violations should be labeled as **Fail**. Borderline or unclear cases are passed to *Expert Adjudication*.

### S.1 Directional Flipping

#### Step 1: Category Check (Original Only)

A sample belongs to the Flipping category if:

- The original title or summary contains at least one *directional or polarity-bearing expression* about performance, outlook, or market reaction.
- It is meaningful to reverse the direction or sentiment while keeping all factual details (numbers, entities, dates, events) unchanged.

Mark **Fail** for the Flipping category only if:

- The original text contains no directional meaning, or
- Another category (pure Numerical or pure Causal) is clearly more appropriate.

**Directional Keyword Families (for non-Expert Annotators).** Annotators may use the following lists to detect flippable directional language. All inflected forms count.

#### 1. Price and performance moves (Up vs. Down).

- Upward vs. downward core moves:  
rise / rises / rose / rising ↔ fall / falls / fell / falling  
climb / climbed / climbing ↔ drop / dropped / dropping  
gain / gains / gained ↔ lose / losses / decline  
jump / jumped / jumping ↔ plunge / plunged / plunging  
soar / soared / soaring ↔ slump / slumped / slumping  
surge / surged / surging ↔ tumble / tumbled / tumbling  
rally / rallied / rallying ↔ retreat / retreating / slip / slipped.

- Acceleration vs. slowdown:  
accelerate / accelerated / accelerating ↔ decelerate / decelerated / slowing  
strengthen / strengthened / strengthening ↔ weaken / weakened / weakening  
speed up ↔ slow down.
- Recovery vs. weakening:  
rebound / rebounded / rebounding ↔ slip / slipped / slipping  
recover / recovered / recovering ↔ deteriorate / deteriorating  
improve / improving ↔ soften / softening.
- External support vs. pressure:  
boost / boosted / boosting ↔ weigh on / drag / pressure  
support / supported ↔ hurt / undermine.

#### 2. Results vs. expectations.

- Beat vs. miss: beat / beats / beating ↔ miss / misses / missed  
top / topped ↔ lag / lags / lagged  
exceeded expectations ↔ fell short of expectations  
above expectations ↔ below expectations.

- Other common phrasings:  
stronger-than-expected ↔ weaker-than-expected  
ahead of estimates ↔ below estimates  
beats consensus ↔ misses consensus  
surpassed forecasts ↔ fell short of forecasts.

### 3. Guidance, ratings, and outlook.

- Guidance revision:  
raise guidance / lifted guidance ↔ cut guidance / slashed guidance  
strong guidance ↔ weak guidance  
raised estimates ↔ trimmed estimates.
- Analyst actions:  
upgrade / upgraded ↔ downgrade / downgraded  
bullish ↔ bearish.
- Outlook tone:  
optimistic outlook ↔ pessimistic outlook  
positive outlook ↔ negative outlook  
robust outlook ↔ soft outlook  
upbeat guidance ↔ downbeat guidance.

### 4. Qualitative sentiment and interpretation.

- Strength vs. weakness: strong / strong performance / strength ↔ weak / weak performance / weakness  
solid ↔ soft / fragile  
robust ↔ shaky / fragile  
resilient ↔ vulnerable / weak  
stable ↔ volatile.
- Positive vs. negative evaluation:  
impressive ↔ underwhelming  
encouraging ↔ disappointing  
notable / significant / substantial ↔ limited / marginal / insignificant.
- Optimism vs. pessimism:  
optimism / optimistic ↔ pessimism / pessimistic  
enthusiasm ↔ fear / concern  
confidence ↔ concern / caution.

- Market reaction:  
investor confidence ↔ investor concern  
upbeat tone ↔ cautious tone  
bullish tone ↔ bearish tone  
sentiment improved ↔ sentiment deteriorated  
market reacted positively ↔ market sold off  
risk-on ↔ risk-off.

### 5. Combined flippable phrases and percentages.

- strong quarter ↔ weak quarter underwhelming guidance ↔ impressive guidance fell short ↔ exceeded solid results ↔ disappointing results robust demand ↔ soft/weak demand improved margins ↔ contracting margins positive sentiment ↔ negative sentiment.
- Any explicit percentage (pattern like “3%”, “12.5%”, “0.4%”) that clearly indicates direction (rise/fall) supports a Flipping interpretation.

#### Step 2: Rewrite Check (Original + Misinformation)

Label the Flipping rewrite as **Pass** only if:

- All factual elements (numbers, companies, tickers, dates, quarters, concrete events) are unchanged.
- Direction or polarity of the key statements is clearly reversed (positive → negative, or negative → positive).
- No new facts, events, or explanations are introduced.
- The new story is logically coherent and financially plausible (no contradictions or impossible behaviour).

If any of these are violated, mark the rewrite **Fail**.

#### S.2 Numerical Perturbation

##### Step 1: Category Check (Original Only)

A sample belongs to Numerical Perturbation if:

- The main information relies on explicit numeric values: percentages, dollar amounts, EPS, margins, costs, user counts, volumes, market share, guidance numbers, etc.

- It is possible to change only the numbers while keeping the narrative direction (up vs. down), entities, and events the same.

Mark the category as **Fail** if:

- Numbers are minor and the main manipulation is directional (Flipping) or causal (Causal Distortion), or
- The news does not meaningfully depend on quantitative details.

## Step 2: Rewrite Check (Original + Misinformation)

A numerical rewrite is **Pass** only if all conditions below hold.

### 1. Scope of change.

- All numeric expressions are modified (percentages, currency amounts, EPS, margins, units, deliveries, user counts, market share, ratios, guidance numbers, etc.).
- No *new* numeric dimensions are introduced (no new subscriber counts, new headcount numbers, new price targets, etc.).

### 2. Direction preserved.

- Increases remain increases; declines remain declines.
- Profit stays profit; loss stays loss.
- A “beat” remains a beat; a “miss” remains a miss.

### 3. Identifiers unchanged.

- Do *not* change time references: years, months, quarters, specific dates.
- Do *not* change identifiers that look numeric but are not financial metrics, such as: product versions (“iPhone 17”), rankings (“Top 10 stocks”), index names.

### 4. Financial plausibility.

- No decline above 100%.
- Market share, engagement rate, and similar ratios must stay between 0% and 100%.
- No impossible margins or EPS (for example, margin above 100%, absurdly large negative values, etc.).

**5. Scenario types for percentages (for annotators).** Annotators should recognise the main scenarios where percentages appear:

- **Scenario A – Stock price movement:** percent change in a stock over a period (“shares rose 12%”, “stock is down 8% YTD”).
- **Scenario B – Company financial metrics:** revenue, profit, EPS, margins, operating costs (“revenue grew 45% YoY”, “gross margin declined 6%”).
- **Scenario C – Operational or user metrics:** units delivered, subscribers, users, shipments, capacity, utilization (“DAUs increased 30%”, “shipments fell 14%”).
- **Scenario D – Market share and ratios:** market share, customer mix, engagement rate, ad load (“market share expanded from 10% to 14%”).
- **Scenario E – Guidance and forecasts:** expected growth, revised guidance, bookings changes (“raised FY revenue guidance by 9%”).
- **Scenario F – Sector or index moves:** index or ETF changes, sector rotation, fund flows (“semiconductor index jumped 18% this month”).

**6. Allowed perturbation ranges (stock-like and company metrics).** For company and stock-like percentages  $X\%$ :

- **$X\%$  up (increase).**
  - If  $X < 130\%$ : allowed either between  $1.5 \times X\%$  and  $200\%$  of  $X$ , or below  $0.5 \times X\%$ .
  - If  $X \geq 130\%$ : allowed either between  $1.5 \times X\%$  and  $2 \times X\%$ , or below  $0.5 \times X\%$ .
- **$X\%$  down (decline).** Declines can never exceed 100%.
  - If  $X < 40\%$ : allowed between  $1.5 \times X\%$  and  $60\%$ , or below  $0.5 \times X\%$ .
  - If  $40\% \leq X < 60\%$ : allowed between  $1.5 \times X\%$  and  $90\%$ , or below  $0.5 \times X\%$ .
  - If  $X \geq 60\%$ : allowed between  $1.3 \times X\%$  and  $100\%$ , or below  $0.5 \times X\%$ .

The same style of ranges applies to revenue growth, EPS changes, margin changes, cost increases/decreases, and operational metrics, with the additional constraint that utilization, engagement, and market share must remain between 0% and 100%.

### 7. Macro-economic percentages (special rules). Macro indicators are checked more strictly:

- **CPI / inflation:** monthly move within about  $\pm 3$  percentage points; annual inflation usually  $\leq 30\%$  unless the original clearly describes hyperinflation.
- **Unemployment:** generally  $\leq 25\%$ ; values above this are allowed only if the article already refers to crisis or youth unemployment.
- **GDP growth:** quarterly change within about  $\pm 10\%$ ; annual change within about  $\pm 20\%$  (excluding explicit crisis rebounds).
- **Policy rates:** central-bank rate decisions usually move by no more than  $\pm 1$  percentage point unless a “shock” move is already mentioned.

Negative macro values are allowed only where historically reasonable (e.g., negative rates in Japan/Eurozone, mild deflation, negative GDP during recessions). If a rewritten number breaks these constraints, annotators must mark **Fail**.

### 8. Coherence checks (all numerical cases).

- Narrative direction is unchanged (up stays up, down stays down).
- Related numbers move consistently (all stronger or all weaker, not contradictory).
- No new causal explanations are introduced solely to justify new numbers.

Any violation of the above implies **Fail** for the numerical rewrite.

## S.3 Sentiment Amplification

### Step 1: Category Check (Original Only)

The sample can be assigned to Sentiment Amplification only if **all** three conditions A, B, and C are satisfied.

**Condition A – Financial KPI present (mandatory).** The article must mention at least one core financial or operational metric, for example:

- revenue, sales
- EPS, earnings, profit, net income
- gross margin, operating margin, margin
- guidance, forecast, outlook
- price target, analyst rating, upgrade/downgrade
- subscribers, users, MAU, DAU
- orders, shipments, bookings, deliveries
- cash flow, free cash flow
- cost, expense, opex, capex
- growth (year-over-year, quarter-over-quarter)
- same-store sales or comparable sales.

If none of these appear, do *not* label the sample as Sentiment Amplification.

**Condition B – Original text is not already extreme.** The original wording should not already use very strong or dramatic emotional language such as:

- crash, plunge, meltdown, collapse
- skyrocket, surge, explode
- disaster, catastrophic
- record high, unprecedented
- panic, bloodbath.

If such words are present, the article is already strongly emotional and is *not* suitable for Sentiment Amplification.

**Condition C – Contains “amplifiable” language.** The text must include uncertain or mild expressions that can be safely strengthened:

- **Hedges / modality (uncertainty):** may, might, could, possibly, potentially, seems, appears, likely, unlikely, expected to, set to, poised to, suggests, indicates, forecast, projected, plans to, aims to, considering.

- **Mild sentiment / weak tone:** modest, limited, slight, somewhat, muted, tepid, lukewarm, soft, stable, steady, resilient, mixed, headwinds, tailwinds, uncertainty, pressures.

If at least one item from this list appears, Condition C is satisfied.

## Step 2: Rewrite Check (Original + Misinformation)

Mark the Sentiment rewrite as **Pass** only if all checks below are satisfied.

### 1. Factual integrity.

- All entities (companies, tickers, executives, institutions) remain identical; no new entities are added.
- All numbers (EPS, revenue, percentages, prices), dates, quarters, and concrete events remain exactly the same.
- Outcomes such as beat/miss, guidance raised/cut, deal announced, investigation started, etc. are unchanged.

Any change to facts, numbers, entities, or events ⇒ **Fail**.

### 2. Allowed tone amplification.

- Neutral or mild wording can be strengthened into more vivid, but still professional, financial-news language: “stock strength” → “remarkable strength”; “rally” → “decisive rally”; “cost pressures” → “strong cost pressures”.
- The rewrite may use a small number (1–3) of stronger adjectives/adverbs.
- Style must remain journalistic and analytical, not advertising.

**3. Over-amplification (must mark Fail).** Label the rewrite **Fail** if it uses sensational or stacked emotional language, for example (non-exhaustive list):

- explosive, stunning, dramatic, electrified, overwhelming, severe, devastating, collapsed, dark cloud, fever pitch, race against time, powerhouse event, explosive surge, deeply troubling.
- The text reads like marketing copy or tabloid drama rather than normal financial reporting.

### 4. Hedges and modality.

- It is acceptable to slightly tighten hedges (for example, “may be worthwhile to compare” → “a closer comparison provides more clarity”), as long as the result remains analytic rather than emotional.
- Do not turn a hedged statement into a strong, dramatic prediction.

### 5. Causality and investor reactions.

- Do *not* introduce new causal chains or investor reactions that were not present in the original: “fueling anxiety”, “raising deep concern”, “casting a shadow of deep skepticism”, “leaving investors on the edge of their seats”, “triggering a race against time”, etc.
- Mild rephrasing of existing relations is fine (for example, “below estimates” → “missed expectations slightly”).

### 6. Summary table for annotators.

| Dimension               | Compliant (Pass)                                 | Non-compliant (Fail)                                               |
|-------------------------|--------------------------------------------------|--------------------------------------------------------------------|
| Facts / numbers / dates | All unchanged                                    | Any fact, number, or date changed or added                         |
| Entities / events       | No new entities or events                        | New companies, deals, crises introduced                            |
| Tone strength           | Mild → moderate (e.g., “remarkable”, “decisive”) | Extreme language (“explosive”, “devastating”, “fever pitch”, etc.) |
| Hedges                  | Slightly reduced, still neutral/analytic         | Turned into absolute dramatic claims                               |
| Causality & reactions   | No new causes or emotions                        | New investor anxiety, panic, race against time, etc.               |

Only when all dimensions are compliant should the rewrite be labeled **Pass** for Sentiment Amplification.

### S.4 Causal Distortion

Causal Distortion candidates are selected when the news text contains an explicit *cause–effect* structure or a clearly stated causal explanation. Candidates are identified using three groups of lexical patterns: (1) causal connectors, (2) catalyst or event terms,

and (3) market–reaction vocabulary. A sample enters the causal–rewriting pipeline if it exhibits at least one such cue and sufficient financial context.

### 1. Causal Connectors

Explicit markers of cause and effect include:

- because, because of;
- since, given that, considering that, in that, insofar as, inasmuch as;
- due to, owing to, on account of, as a result of, in light of, in view of, thanks to;
- therefore, thus, hence, consequently, accordingly, as a result, as a consequence, in consequence, thereby;
- result in, results in, resulted in, resulting in;
- lead to, leads to, led to, leading to;
- cause, causes, caused, causing;
- bring about, brings about, brought about, bringing about;
- end up in, end up with, ends up in, ended up in;
- trigger, triggers, triggered, triggering.

### 2. Catalyst and Event Terms

These denote events that frequently act as “causes” in financial news:

- product recall, recall;
- probe, investigation;
- lawsuit, litigation;
- regulatory action, fine, penalty, sanctions;
- guidance cut, guidance raise, lifted guidance, slashed guidance, downgrade, upgrade, price target, outlook, forecast;
- outage, breach, hack, cyberattack;
- strike, walkout, layoff, layoffs;
- acquisition, merger, deal;
- supply shortage, supply disruption, supply glut;
- delay, postpone, halt, suspend;
- defect, safety issue, recertification;
- earnings miss, earnings beat, EPS, margins.

### 3. Market Reaction Vocabulary

To capture event–reaction narratives, we additionally consider:

- shares rise, shares rose, rally, rallied, jump, jumped, surge, surged, gain, gained;
- shares fall, fell, drop, dropped, slump, slumped, plunge, plunged, sink, slide, tumble, tumbled;
- stock spike, plunge, soar, slump, jump, drop;
- sell-off, rout, rally;
- market rebound, market sell-off, market turmoil, volatility.

### Causal Candidate Criterion

A news item is treated as a CAUSAL DISTORTION CANDIDATE if it contains:

1. at least one causal connector or catalyst pattern; **or**
2. a clearly expressed event–event causal explanation, even when no explicit connective appears.

Such candidates are subsequently evaluated through the full causal–rewriting guideline (category verification and distortion validation).

### T Task Prompts

This appendix lists the exact prompts used for all experiments, including zero-shot and few-shot settings for Task 1 and Task 2.

#### T.1 Task 1: Reference-free Detection Prompts

##### Zero-shot Prompt

You are a financial misinformation detector.  
Please check whether the following information is false and output the answer [true/false].  
Return only one label.

##### Two-shot Prompts by Manipulation Type Flipping (2-shot)

You are a financial misinformation detector.  
Please check whether the following information is false and output the answer [true/false].  
Return only one label.

Example 1:  
Input:

Li Auto Stock Falls After Wall Street Downgrade. It's About China. JPMorgan analyst Nick Lai cut his rating on Li Auto stock to Hold from Buy. He is worried about slowing demand.

Output:  
true

Example 2:

Input:  
Li Auto stock jumps after Wall Street upgrade. It's about China. JPMorgan analyst Nick Lai raised his rating on Li Auto stock to Buy from Hold. He is optimistic about accelerating demand.

Output:  
false

### **Numerical (2-shot)**

You are a financial misinformation detector. Please check whether the following information is false and output the answer [true/false]. Return only one label.

Example 1:

Input:  
Alibaba to raise \$3.2 billion via convertible bond to fund cloud growth. Chinese e-commerce leader Alibaba said on Thursday it plans to raise \$3.2 billion through the sale of a zero-coupon convertible bond to fund international expansion and strengthen cloud computing. The bond will be the largest of its kind this year, showed Dealogic data, eclipsing DoorDash's \$2.75 billion deal in May. Alibaba said it would use nearly 80% of the proceeds to expand data centres, upgrade technology and improve services to meet demand for cloud solutions.

Output:  
true

Example 2:

Input:  
Alibaba to raise \$5.1 billion via convertible bond to fund cloud growth. Chinese e-commerce leader Alibaba said on Thursday it plans to raise \$5.1 billion through the sale of a zero-coupon convertible bond to fund international expansion and strengthen cloud computing. The bond will be the largest of its kind this year, according to Dealogic data, surpassing DoorDash's \$4.3 billion deal in May. Alibaba stated it would use nearly 35%

of the proceeds to expand data centres, upgrade technology, and improve services to meet demand for cloud solutions.

Output:  
false

### **Sentiment (2-shot)**

You are a financial misinformation detector. Please check whether the following information is false and output the answer [true/false]. Return only one label.

Example 1:

Input:  
Glaukos Builds Dual Growth Engines Amid Market Headwinds. GKOS raises revenue guidance after record 1H sales, fueled by iDose TR adoption and expanding international glaucoma momentum.

Output:  
true

Example 2:

Input:  
Glaukos demonstrates impressive resilience by building dual growth engines despite market headwinds, as GKOS lifts revenue guidance following record first-half sales, powered by strong iDose TR adoption and accelerating international glaucoma momentum.

Output:  
false

### **Causal (2-shot)**

You are a financial misinformation detector. Please check whether the following information is false and output the answer [true/false]. Return only one label.

Example 1:

Input:  
Walmart Stock Up 25% in 2025: What's the Smart Move for 2026? WMT's 2025 rally, fueled by e-commerce strength and higher-margin growth, sets the stage for a balanced but watchful outlook heading into 2026.

Output:  
true

Example 2:

Input:  
Walmart Stock Up 25% in 2025: What's the Smart Move for 2026? WMT's 2025 rally, attributed to strategic cost management and operational efficiencies, sets the

stage for a balanced but watchful outlook heading into 2026.

Output:  
false

### **Eight-shot Prompt**

You are a financial misinformation detector. Please check whether the following information is false and output the answer [true/false]. Return only one label.

Example 1:

Input:  
Alibaba to raise \$3.2 billion via convertible bond to fund cloud growth.  
Chinese e-commerce leader Alibaba said on Thursday it plans to raise \$3.2 billion through the sale of a zero-coupon convertible bond to fund international expansion and strengthen cloud computing. The bond will be the largest of its kind this year, showed Dealogic data, eclipsing DoorDash's \$2.75 billion deal in May.  
Alibaba said it would use nearly 80% of the proceeds to expand data centres, upgrade technology and improve services to meet demand for cloud solutions.

Output:  
true

Example 2:

Input:  
Alibaba to raise \$5.1 billion via convertible bond to fund cloud growth.  
Chinese e-commerce leader Alibaba said on Thursday it plans to raise \$5.1 billion through the sale of a zero-coupon convertible bond to fund international expansion and strengthen cloud computing. The bond will be the largest of its kind this year, according to Dealogic data, surpassing DoorDash's \$4.3 billion deal in May.  
Alibaba stated it would use nearly 35% of the proceeds to expand data centres, upgrade technology, and improve services to meet demand for cloud solutions.

Output:  
false

Example 3:

Input:  
Glaukos Builds Dual Growth Engines Amid Market Headwinds.  
GKOS raises revenue guidance after record 1H sales, fueled by iDose TR adoption and expanding international glaucoma momentum.

Output:  
true

Example 4:

Input:  
Glaukos demonstrates impressive resilience by building dual growth engines despite market headwinds, as GKOS lifts revenue guidance following record first-half sales, powered by strong iDose TR adoption and accelerating international glaucoma momentum.

Output:  
false

Example 5:

Input:  
Li Auto Stock Falls After Wall Street Downgrade. It's About China.  
JPMorgan analyst Nick Lai cut his rating on Li Auto stock to Hold from Buy. He is worried about slowing demand.

Output:  
true

Example 6:

Input:  
Li Auto stock jumps after Wall Street upgrade. It's about China.  
JPMorgan analyst Nick Lai raised his rating on Li Auto stock to Buy from Hold.  
He is optimistic about accelerating demand.

Output:  
false

Example 7:

Input:  
Walmart Stock Up 25% in 2025: What's the Smart Move for 2026?  
WMT's 2025 rally, fueled by e-commerce strength and higher-margin growth, sets the stage for a balanced but watchful outlook heading into 2026.

Output:  
true

Example 8:

Input:  
Walmart Stock Up 25% in 2025: What's the Smart Move for 2026?  
WMT's 2025 rally, attributed to strategic cost management and operational efficiencies, sets the stage for a balanced but watchful outlook heading into 2026.

Output:  
false

## **T.2 Task 2: Comparative Diagnosis Prompts**

### **Zero-shot Prompt**

You are a financial misinformation type detector.

Given a pair of original news and misinformation, identify the misinformation type:

[numerical / flipping / sentiment / causal].

Type Definitions:

Numerical: Alters quantitative facts while keeping the narrative structure.

Flipping: Reverses polarity or evaluation while preserving factual content.

Sentiment: Changes emotional tone or intensity without altering facts or numbers.

Causal: Adds or modifies cause effect relationships or explanations.

Return only one label.