

The 2nd Workshop on Misinformation Detection in the Era of LLMs (MisD) shared task: Reference Free Financial Misinformation Detection

Yuechen Jiang¹, Yuyan Wang¹, Tianlei Zhu², Peter Carragher³, Yixiang Zheng¹,
Zhiwei Liu¹, Yupeng Cao⁴, Jimin Huang^{1,5}, Sophia Ananiadou^{1,6}

¹University of Manchester, UK ²Columbia University, US ³Carnegie Mellon University, US

⁴Stevens Institute of Technology, US ⁵The Fin AI, US ⁶ELLIS Manchester

{yuechen.jiang, yixiang.zheng, jimin.huang}@postgrad.manchester.ac.uk,

{yuyan.wang-2, zhiwei.liu, sophia.ananiadou}@manchester.ac.uk

tz2617@columbia.edu, petercarragher@cmu.edu, ycao33@stevens.edu

Abstract

We present an overview of the Reference-Free Financial Misinformation Detection shared task at the 2nd Workshop on Misinformation Detection in the Era of LLMs (MisD). Built on RFC-Bench, the task requires systems to classify a single financial news paragraph as factual or minimally manipulated, without access to any external evidence or source document. This reference-free setting isolates the difficulty of reasoning about internal semantic consistency and financial plausibility. The task attracted multiple participating teams whose approaches span three paradigms: parameter-efficient LLM adaptation, discriminative learning with controlled augmentation, and reasoning-augmented multi-agent frameworks. On the private test set, the top system reaches an F1 of 0.9629, while a 2-shot GPT-4.1 baseline remains close to chance at 0.5686. The results show clear progress on pattern-level detection, but also indicate that reference-free reasoning from a single paragraph remains an open problem.

Introduction

The rapid development of large language models (LLMs) has intensified interest in financial misinformation detection, particularly as these models are increasingly deployed in high-stakes decision-making scenarios. Existing research has advanced through a range of benchmarks and shared tasks that evaluate models under diverse conditions, including multilingual variation and scenario-induced bias (Liu et al. 2026a), as well as fine-grained span detection and reasoning-oriented inference (Liu et al. 2026c,b). In parallel, community-driven initiatives such as the FinNLP shared task (Liu et al. 2025b) and the MisD workshop series (Liu et al. 2025a) have contributed standardized datasets and evaluation protocols that support systematic comparison. More broadly, misinformation detection has been explored across multiple paradigms, including deceptive language identification (Mihalcea and Strapparava 2009), fake news classification (Wang 2017), and evidence-based verification (Thorne et al. 2018). Additional efforts have incorporated stance detection and social context modeling (Hanselowski et al. 2018; Shu et al. 2018), as well as domain-specific datasets in health and scientific misinformation (Dai, Sun,

and Wang 2020; Cui and Lee 2020). Despite this progress, most existing approaches rely on external evidence, retrieved documents, or source claims, leaving underexplored the more challenging setting in which a model must assess the credibility of a single paragraph in isolation.

This limitation is particularly critical in financial contexts. Financial narratives can be subtly manipulated through minimal edits that preserve fluency while altering meaning, such as modifying numerical values, reversing directional implications, or introducing misleading causal relationships. Prior work has shown that even small perturbations in financial text can significantly influence downstream systems, including stock prediction models (Xie et al. 2022). At the same time, recent financial misinformation benchmarks have focused on multimodal reasoning and explainability (Rangapur et al. 2023), or on verification over complex and long-form financial documents (Zhao et al. 2024). However, these settings continue to depend on external grounding. Meanwhile, LLMs are known to generate coherent yet factually inconsistent content, exhibiting hallucination and reliance on surface-level plausibility (Ji et al. 2023; Alansari and Luqman 2025). Methods based on detecting factual inconsistencies have also highlighted the importance of internal coherence signals (Gupta et al. 2025). Consequently, detecting misinformation without external references requires models to reason about semantic consistency, numerical plausibility, and domain-specific financial logic, rather than relying on retrieval or evidence matching.

To address this gap, we present the Reference-Free Financial Misinformation Detection shared task at the 2nd Workshop on Misinformation Detection in the Era of LLMs (MisD). The task requires systems to classify a single financial news paragraph as factual or minimally manipulated, without access to any external evidence or original source. This formulation isolates the core challenge of reference-free reasoning, where models must detect subtle semantic inconsistencies and plausibility violations from local textual signals alone. Participating systems explore diverse approaches, including parameter-efficient LLM adaptation, discriminative learning with controlled augmentation, and reasoning-augmented multi-agent frameworks.

Our task is built upon RFC-Bench (Jiang et al. 2026), which demonstrates that removing access to reference in-

formation leads to near-chance performance even for strong models. This finding suggests that current systems are not limited by domain knowledge alone, but by their inability to reliably reason about belief validity from a single paragraph. The shared task therefore provides a controlled setting to evaluate progress on this problem, and to better understand the extent to which current models can move beyond pattern recognition toward genuine reference-free reasoning.

Task and Dataset

Task Definition

We adopt the reference-free financial misinformation detection task (Task 1) introduced in Jiang et al. (2026). The task is formulated as a binary classification problem at the paragraph level.

Given an input financial news paragraph x , the model is required to predict whether it is:

- **Factual:** an original, unaltered paragraph, or
- **Manipulated:** a minimally perturbed version that preserves fluency but alters the underlying meaning.

Formally, the task can be written as:

$$y = f(x), \quad y \in \{\text{Factual, Manipulated}\}. \quad (1)$$

A key characteristic of this task is that it is **reference-free**: models must make decisions without access to external evidence, source documents, or the original paragraph. This setting requires models to rely on internal reasoning about semantic consistency, discourse coherence, and financial plausibility, making it substantially more challenging than standard fact-checking or retrieval-based verification tasks.

Dataset

The dataset used in this shared task is directly based on RFC-Bench (Jiang et al. 2026). It consists of paragraph-level financial news data constructed from real-world sources.

The dataset contains approximately **1.8K paragraph pairs**, where each pair includes:

- an original (factual) paragraph, and
- a corresponding manipulated (counterfactual) version.

Each manipulated instance is generated through controlled, minimal edits that preserve surface fluency while introducing subtle semantic shifts. The dataset covers multiple types of financial narrative manipulations, including directional changes, numerical perturbations, sentiment shifts, and causal distortions, as described in Jiang et al. (2026).

For the purpose of this shared task, only the **reference-free detection setting** is used. Models are evaluated on their ability to distinguish factual and manipulated paragraphs without any additional context or grounding information.

Participants and Automatic Evaluation

Participants

A total of multiple teams participated in the shared task, submitting systems for the reference-free financial misinformation detection task. Participants were allowed to use

any modeling approach, including large language models (LLMs), fine-tuned classifiers, or hybrid systems.

Submissions were evaluated through an online leaderboard system. Participants could submit their predictions multiple times and receive real-time feedback on the public test set, enabling iterative development and tuning.

Data Splits and Evaluation Protocol

The dataset is divided into three subsets:

- **Training set:** 2,000 instances (1,000 factual-manipulated pairs), publicly available.
- **Public test set:** 652 instances, used for leaderboard feedback during the competition.
- **Private test set:** 1,000 instances (500 pairs), used for final ranking.

During the competition, participants were allowed to submit predictions on the test server and observe their performance on the **public test set**. The **final ranking**, however, was determined exclusively based on performance on the **private test set**, ensuring a fair and unbiased evaluation.

Evaluation Metrics

Following Jiang et al. (2026), we adopt standard binary classification metrics to evaluate system performance, including Accuracy, Precision, Recall, and F1 score. Accuracy measures the overall proportion of correctly classified instances, while Precision and Recall capture the trade-off between false positives and false negatives, respectively. The F1 score, defined as the harmonic mean of Precision and Recall, provides a balanced assessment of model performance under class imbalance and is widely used in misinformation detection tasks.

In this shared task, **F1 score is used as the primary ranking metric**. This choice is motivated by the nature of the task: correctly identifying manipulated instances without over-predicting either class is critical, and F1 score explicitly captures this balance. While Accuracy provides a general overview of performance, it may obscure asymmetric errors, making Precision, Recall, and F1 more informative for comparing systems in this setting.

Results and Analysis

Table 1 presents the final results on the private test set, where systems are ranked by F1 score. The top-performing teams achieve strong and consistent performance, with Fact4ac reaching an F1 score of 0.9629, followed by AISper (0.9589) and Coherence (0.9549), indicating effective detection of manipulated financial narratives and a balanced trade-off between Precision and Recall. In contrast, baseline approaches remain close to chance-level performance, with a 2-shot GPT-4.1 baseline achieving 0.5686 and mfPE reaching 0.5547. This substantial gap highlights the difficulty of the reference-free setting, where models must rely solely on internal textual signals without access to external evidence. Overall, the results suggest that while advanced methods that leverage structured reasoning, data augmentation, or

Team	Accuracy	Precision	Recall	F1 Score
Fact4ac	0.963	0.9654	0.9626	0.9629
AIspers	0.959	0.9619	0.9586	0.9589
Coherence	0.955	0.9566	0.9547	0.9549
DeepTruth	0.755	0.7551	0.7551	0.7550
Baseline (GPT-4.1, 2-shot)	0.570	0.5701	0.5694	0.5686
mfPE	0.555	0.5556	0.5554	0.5547

Table 1: Leaderboard on the private test set. Systems are ranked by F1 score.

parameter-efficient adaptation can significantly improve performance, reference-free financial misinformation detection remains a challenging problem that requires deeper modeling of semantic consistency and financial plausibility.

Methods of Each Team

In this section, we summarize the representative approaches adopted by participating teams. Despite sharing the same reference-free setting, the submitted systems exhibit diverse methodological designs, ranging from prompt-based LLM adaptation to hybrid architectures combining reasoning and discriminative modeling.

Fact4ac (Rank 1)

The top-ranked team, Fact4ac, adopts a hybrid strategy combining **in-context learning** and **parameter-efficient fine-tuning**. Their approach integrates few-shot prompting with Low-Rank Adaptation (LoRA) to adapt large language models to the subtle patterns of financial misinformation. Specifically, the system constructs a unified prompt template consisting of task instructions and representative examples (one factual and one manipulated), enabling the model to infer implicit patterns of semantic perturbation. This prompting strategy is further strengthened by PEFT-based fine-tuning, allowing the model to internalize domain-specific cues without incurring high computational cost. This combination of **few-shot reasoning and lightweight fine-tuning** proves highly effective, achieving the best overall performance.

AIspers (Rank 2)

AIspers proposes a **two-stage framework** centered on robust representation learning and data augmentation. The key innovation lies in introducing **benign controls** alongside perturbations to prevent models from learning superficial “edited-text” shortcuts. Their pipeline includes:

- LLM-guided generation of both label-changing perturbations and label-preserving rewrites
- Construction of **semantic groups** to enforce group-aware training
- A multi-head RoBERTa-based classifier trained with auxiliary objectives, including contrastive learning and ranking loss

Additionally, an optional LLM-based reranker is applied to uncertain cases during inference. This approach emphasizes **distribution robustness and fine-grained semantic discrimination**, leading to strong generalization performance.

Coherence (Rank 3)

The Coherence team introduces a **courtroom-inspired framework** to address the well-known **LLM compliance bias**. For each input paragraph, the system generates:

- a **prosecution argument** (why the claim is false), and
- a **defense argument** (why the claim may be true)

These adversarial perspectives are then fed into a multi-channel neural classifier based on ModernBERT, combined with CNN, BiGRU, and attention layers for final judgment. By explicitly modeling **conflicting reasoning paths**, this method improves the model’s ability to detect subtle inconsistencies in financial narratives.

DeepTruth (Rank 4)

DeepTruth focuses on incorporating **explicit reasoning signals** into the classification process through a rationale-guided framework. The method first uses an LLM (DeepSeek-V3) to generate two complementary rationales:

- a **text-based rationale** capturing semantic consistency, and
- a **commonsense rationale** focusing on financial plausibility

These rationales are then combined with the original input and fed into a BERT-based classifier with a co-attention mechanism. This approach highlights the importance of **interpretable reasoning augmentation**, although its performance remains below the top systems.

mfPE

The mfPE system adopts a **multi-agent prompt engineering framework**. It decomposes the task into multiple perspectives by generating:

- linguistic feature analysis reports,
- fact-checking reports, and
- (simulated) comment analysis reports

These reports are aggregated and used to construct improved prompts through an **iterative prompt optimization process**, where prompts are refined based on historical performance. The final prediction is produced by an LLM conditioned on both the reports and the optimized prompt.

Summary

Overall, the submitted systems can be broadly categorized into three paradigms:

- **LLM adaptation approaches** (e.g., Fact4ac), focusing on prompting and efficient fine-tuning
- **Discriminative learning with augmentation** (e.g., AIsper), emphasizing robustness and representation learning
- **Reasoning-augmented frameworks** (e.g., Coherence, DeepTruth, mfPE), introducing explicit reasoning signals via multi-agent or multi-perspective designs

The strong performance of top systems suggests that **combining structured reasoning signals with robust training objectives** is key to addressing the challenges of reference-free financial misinformation detection.

Discussion

The shared task results show a large performance gap between top systems and baseline approaches, with F1 improving from 0.5686 (GPT-4.1, 2-shot) to above 0.95. This improvement is not due to stronger raw reasoning alone, but primarily to better **structural modeling of the task**. The top systems introduce intermediate representations that reshape the problem: Fact4ac combines few-shot prompting with parameter-efficient fine-tuning to adapt to subtle manipulation patterns; AIsper leverages benign controls, semantic grouping, and ranking objectives to prevent shortcut learning; Coherence introduces adversarial reasoning via prosecution and defense arguments to mitigate LLM compliance bias. In contrast, the baseline relies on direct single-pass classification without additional structure, which leads to near-chance performance.

The difficulty of the task is consistent with findings in RFC-Bench. When models are given only a single paragraph, performance remains near chance, whereas providing the original paragraph for contrast leads to dramatic improvements. This indicates that the bottleneck lies in **reference-free reasoning**, rather than domain knowledge alone. The dataset is specifically designed to preserve surface plausibility while altering the implied meaning through minimal edits (e.g., numerical, directional, or causal changes), making superficial cues ineffective. As a result, models must determine whether a paragraph should be trusted based solely on internal consistency and plausibility, which is fundamentally more difficult than traditional fact-checking.

Despite strong leaderboard performance, the results suggest that participants have **not fully solved reference-free reasoning**. Instead, most systems rely on **distribution-aware detection mechanisms**. For example, AIsper explicitly frames the task as identifying whether a paragraph resembles a benchmark-real instance rather than verifying truth against the real world, and still shows weaknesses under compositional or unseen perturbations. Similarly, Fact4ac emphasizes that pure logical parsing without adaptation is insufficient, and instead relies on fine-tuning to capture dataset-specific patterns. These observations indicate that models are learning to detect *benchmark-*

consistent perturbations, rather than performing general reasoning about belief validity.

Moreover, even reasoning-augmented systems depend on auxiliary scaffolding. Coherence improves performance by generating opposing arguments, while DeepTruth introduces textual and commonsense rationales before classification. These approaches enhance performance by restructuring the input rather than demonstrating that models can directly reason from the paragraph itself. In this sense, current methods reduce the difficulty of the task by introducing implicit contrast or additional signals, rather than solving the underlying reasoning problem.

Overall, the shared task highlights a key distinction: current models are becoming effective at **detecting suspicious patterns without references**, but still struggle to **decide what to believe from a single paragraph**. Progress on this benchmark, therefore, reflects improvements in structured detection and representation learning, rather than a complete solution to reference-free reasoning. Future work should focus on generalization beyond controlled perturbations, robustness to unseen manipulation types, and mechanisms for uncertainty-aware decision making when evidence is inherently limited.

Conclusion

We present the Reference-Free Financial Misinformation Detection shared task at the 2nd Workshop on Misinformation Detection in the Era of LLMs (MisD), which introduces a challenging evaluation setting requiring models to assess the credibility of financial news paragraphs without external evidence. The results demonstrate that while top-performing systems achieve strong performance through structured modeling strategies such as parameter-efficient adaptation, controlled augmentation, and reasoning-augmented frameworks, baseline approaches remain near chance level. This gap highlights that current progress is driven more by task-specific pattern recognition than by genuine reference-free reasoning. Beyond the benchmark itself, this shared task contributes to the research community by establishing a standardized testbed, fostering diverse methodological exploration, and enabling systematic comparison of approaches under a unified setting. The findings emphasize the need for future work on generalization, robustness to unseen perturbations, and deeper modeling of semantic consistency and financial plausibility, ultimately advancing the study of reasoning capabilities in LLMs.

Acknowledgments

This work was supported by the NVIDIA Academic Grant Program using 32K A100 GPU-hours on Brev. We thank all shared task participants, organizers, and reviewers for their valuable contributions, as well as The Fin AI community for its research support, feedback, and collaborative environment that made this work possible.

References

- Alansari, A.; and Luqman, H. 2025. Large Language Models Hallucination: A Comprehensive Survey. *arXiv preprint arXiv:2510.06265*.
- Cui, L.; and Lee, D. 2020. CoAID: COVID-19 Healthcare Misinformation Dataset. *arXiv preprint arXiv:2006.00885*.
- Dai, E.; Sun, Y.; and Wang, S. 2020. Ginger Cannot Cure Cancer: Battling Fake Health News with a Comprehensive Data Repository. *arXiv preprint arXiv:2002.00837*. Dataset: FakeHealth (HealthStory, HealthRelease); DOI: 10.5281/zenodo.3841644.
- Gupta, R.; Panicker, P. H.; Bhatia, S.; and Ramakrishnan, G. 2025. Consistency Is the Key: Detecting Hallucinations in LLM Generated Text By Checking Inconsistencies About Key Facts. *arXiv preprint arXiv:2511.12236*.
- Hanselowski, A.; PVS, A.; Schiller, B.; Caspelherr, F.; Chaudhuri, D.; Meyer, C. M.; and Gurevych, I. 2018. A Retrospective Analysis of the Fake News Challenge Stance Detection Task. *arXiv preprint arXiv:1806.05180*.
- Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12): 1–38.
- Jiang, Y.; Liu, Z.; Cao, Y.; He, Y.; Xu, Z.; Xu, C.; Deng, Z.; Tiwari, P.; Chen, X.; Lopez-Lira, A.; Huang, J.; Tsujii, J.; and Ananiadou, S. 2026. All That Glisters Is Not Gold: A Benchmark for Reference-Free Counterfactual Financial Misinformation Detection. *arXiv:2601.04160*.
- Liu, Z.; Cao, Y.; Jiang, Y.; Kabir, M.; Giannouris, P.; Xu, C.; Xu, Z.; Zhu, T.; Tariquzzaman, M.; Papadopoulos, T.; Wang, Y.; Qian, L.; Peng, X.; Xie, Z.; Yuan, Y.; Almheiri, S.; Alnajjar, A.; Chen, M.; Stuart, H.; Thompson, P.; Tiwari, P.; Lopez-Lira, A.; Liu, X.; Huang, J.; and Ananiadou, S. 2026a. Same Claim, Different Judgment: Benchmarking Scenario-Induced Bias in Multilingual Financial Misinformation Detection. *arXiv:2601.05403*.
- Liu, Z.; de Kock, C.; Knight, P.; Hovy, E.; and Ananiadou, S. 2025a. The 1st Workshop on Misinformation Detection in the Era of LLMs (MisD 2025). In *Proceedings of the 19th International AAAI Conference on Web and Social Media (ICWSM 2025) Workshops*.
- Liu, Z.; Guo, R.; Qu, B.; Jiang, Y.; Peng, M.; Xie, Q.; and Ananiadou, S. 2026b. RAAR: Retrieval Augmented Agentic Reasoning for Cross-Domain Misinformation Detection. *arXiv:2601.04853*.
- Liu, Z.; Thompson, P.; Rong, J.; Qu, B.; Guo, R.; Peng, M.; Xie, Q.; and Ananiadou, S. 2026c. MisSpans: Fine-Grained False Span Identification in Cross-Domain Fake News. *arXiv:2601.04857*.
- Liu, Z.; Wang, K.; Bao, Z.; Zhang, X.; Dong, J.; Yang, K.; Kabir, M.; Giannouris, P.; Xing, R.; Park, S.; Kim, J.; Li, D.; Xie, Q.; and Ananiadou, S. 2025b. FinNLP-FNP-LLMFinLegal-2025 Shared Task: Financial Misinformation Detection Challenge Task. In Chen, C.-C.; Moreno-Sandoval, A.; Huang, J.; Xie, Q.; Ananiadou, S.; and Chen, H.-H., eds., *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)*, 271–276. Abu Dhabi, UAE: Association for Computational Linguistics.
- Mihalcea, R.; and Strapparava, C. 2009. The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. In Su, K.-Y.; Su, J.; Wiebe, J.; and Li, H., eds., *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 309–312. Suntec, Singapore: Association for Computational Linguistics.
- Rangapur, A.; Wang, H.; Jian, L.; and Shu, K. 2023. FINFACT: A Benchmark Dataset for Multimodal Financial Fact Checking and Explanation Generation. *arXiv preprint arXiv:2309.08793*. Version v2, posted 1 May 2024.
- Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 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*.
- Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; and Mittal, A. 2018. FEVER: A Large-Scale Dataset for Fact Extraction and VERification. In *NAACL-HLT 2018*.
- Wang, W. Y. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *ACL 2017 (Short Papers)*.
- Xie, Y.; Wang, D.; Chen, P.-Y.; Xiong, J.; Liu, S.; and Koyejo, S. 2022. A Word is Worth A Thousand Dollars: Adversarial Attack on Tweets Fools Stock Prediction. *arXiv preprint arXiv:2205.01094*.
- Zhao, Y.; Long, Y.; Jiang, Y.; Wang, C.; Chen, W.; Liu, H.; Zhang, Y.; Tang, X.; Zhao, C.; and Cohan, A. 2024. FinDVer: Explainable claim verification over long and hybrid-content financial documents. *arXiv preprint arXiv:2411.05764*.