Ask Asper at the Financial Misinformation Detection Challenge Task: Enhancing Financial Decision-Making: A Dual Approach Using Explainable LLMs for Misinformation Detection

Sonal Singh, Rahul Mehta, Yadunath Gupta, Soudip Roy Chowdhury
Asper.ai Technologies Pvt. Ltd.
{sonal.singh,rahul.mehta2,yadunath.gupta,soudip.chowdhury}@asper.ai

Abstract

The integrity of the market and investor confidence are seriously threatened by the proliferation of financial misinformation via digital media. Existing approaches such as fact check, lineage detection and others have demonstrated significant progress in detecting financial misinformation. In this paper, we present a novel two-stage framework leveraging large language models (LLMs) to identify and explain financial misinformation. The framework first employs a GPT-4 model fine-tuned on financial datasets to classify claims as "True," "False," or "Not Enough Information" by analyzing relevant financial context. To enhance classification reliability, a second LLM serves as a verification layer, examining and refining the initial model's predictions. This dual-model approach ensures greater accuracy in misinformation detection through cross-validation.

Beyond classification, our methodology emphasizes generating clear, concise, and actionable explanations that enable users to understand the reasoning behind each determination. By combining robust misinformation detection with interpretability, our paradigm advances AI system transparency and accountability, providing valuable support to investors, regulators, and financial stakeholders in mitigating misinformation risks.

1 Introduction

The integrity of financial markets faces an unprecedented challenge from the proliferation of misinformation, which fundamentally undermines investor trust and threatens economic stability. Financial misinformation, a particularly harmful subset of deceptive content, can significantly distort investor behavior, market perspectives, and lead to suboptimal financial decisions. This phenomenon manifests in various forms, from fraudulent financial statements to misleading investment advice, carrying severe implications for both individual and institutional stakeholders (Carpenter, 2023). The exponential growth of digital platforms facilitating real-time financial transactions has amplified the impact of such misinformation, necessitating robust detection and mitigation strategies (Chung et al., 2022).

While existing frameworks primarily focus on identifying fraudulent claims, they often lack the transparency necessary to establish user trust. The emergence of advanced artificial intelligence (AI) models, particularly Large Language Models (LLMs), presents promising avenues for detecting and understanding financial misinformation. However, the integration of these technologies with practical financial applications remains an underexplored area, especially concerning explainability and reliability.

This research introduces a novel two-stage methodology that leverages LLMs, enhanced through financial dataset fine-tuning, to classify financial assertions into three categories ("True," "False," or "Not Enough Information") while providing concise, comprehensible explanations for these classifications. Our approach implements a refined GPT-4 model that evaluates the context of financial claims and predicts their veracity, followed by a secondary LLM serving as a "judge" to review and refine initial classifications. This dual-layer verification mechanism enhances reliability in the decision-making process through improved accuracy and comprehensibility (Zheng, 2023).

The next section focuses on the related prior work. In Section 3, we will discuss the proposed architecture, its working, and its advantages. Section 4 will provide an in-depth explanation of the experimental setup and evaluation methodology. Following this, Section 5 will present the results of our experiments, accompanied by a detailed analysis. Finally, we conclude the paper in Section 6, outlining the future work planned to extend this research.

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 283–287
January 19–20, 2025. ©2025 Association for Computational Linguistics

2 Literature Review

Deep learning and natural language processing (NLP) techniques have gained significant attention in detecting financial disinformation and fake news. Numerous models have been proposed, each with unique strengths and limitations.

FNFNet (Xie et al., 2021) employs convolutional neural networks (CNNs) for extracting information from news articles, achieving a remarkable accuracy of 98.46

FMDLlama (Liu et al., 2024), built on Llama3.1 and utilizing the Financial Misinformation Detection Instruction Dataset (FMDID), excels in multi-task learning for classification and explanation generation. Despite its promise, its effectiveness is constrained by limited dataset diversity and the absence of real-world evaluation benchmarks.

Traditional machine learning methods have evolved into deep learning-based approaches like CNNs and LSTMs, which improve classification precision through automated feature extraction (Carpenter, 2023) (Moore et al., 2012). However, these methods often rely heavily on specific datasets, lack generalizability, and require multi-modal data integration.

FinBERT (Yang and Zhang, 2020), a domain-adaptive language model trained on financial texts, captures financial terminology and sentiment effectively. Nonetheless, it struggles to keep up with changing market conditions and financial jargon, underscoring the need for continuous updates.

DFDR (Yang and Liu, 2023) takes a multimodal approach by integrating textual analysis with market data, including trading volumes and real-time signals. While this enhances detection capabilities, it encounters challenges like high computational costs and difficulties in maintaining real-time performance.

The Temporal-Aware Language Model (Zhang and Wang, 2023) focuses on handling time-sensitive financial data by incorporating temporal dependencies and market dynamics. Despite its strength in timely detection, it faces resource constraints and struggles with long-term dependency modeling.

CrossFin (Wang and Liu, 2022) unifies data from diverse sources, including social media, news platforms, and financial streams, enabling effective cross-platform detection. However, its performance consistency across platforms remains a challenge.

Finally, FinGPT (Chen and Zhang, 2023), an open-source financial language model with specialized pre-training, demonstrates a strong ability to understand complex financial narratives. Its main drawbacks include slower inference speeds, optimization issues for model size, and challenges in adapting to rapidly changing market trends.

While models like FNFNet, FMDLlama, and FinBERT have advanced financial misinformation detection, significant gaps remain. These include the need for integrated multimodal approaches, better interpretability, robust benchmarks, and solutions for overfitting and dataset limitations.

3 Proposed Architecture

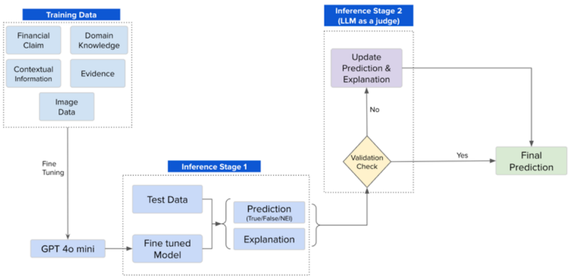

Figure 1: Logical architecture of the proposed solution

Figure 1 demonstrates the logical architecture of our suggested two-step framework.

In the first step, we categorize financial claims as True, False, or Not Enough Information using a fine-tuned GPT-4 model specifically trained on financial data. This model leverages domain-specific financial traits and contextual knowledge, which are provided in the dataset, including claims, justifications, issues, evidence, image URLs, and image content. By incorporating these elements, the model ensures that the classification aligns with accepted financial logic and principles.

To further enhance the reliability and accuracy of the initial classification, we introduce a second layer of verification using the "LLM as a Judge" technique (Zheng, 2023). In this stage, a second instance of GPT-4 serves as an impartial arbiter to assess the accuracy of the first model's predictions. This LLM evaluates the classification's justification, compares it to pertinent financial information, and renders an assessment of the classification's accuracy. If any discrepancies are found, the judge updates the prognosis, providing a thorough justification for the change. This ensures that the final classification benefits from both increased accuracy and a clear, intelligible explanation.

4 Experimentation Setup

4.1 Dataset

We used the FIN-FACT dataset (Rangapur et al., 2024), a comprehensive collection of financial claims spanning domains such as Income, Finance, Economy, Budget, Taxes, and Debt. The dataset categorizes claims into three labels: True, False, and NEI (Not Enough Information), facilitating accurate assessment of financial statements.

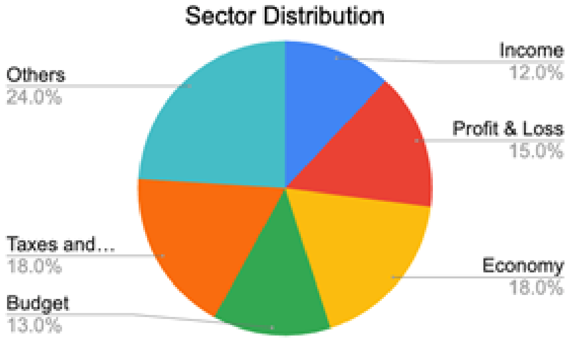

Figure 2: Sector-wise distribution of claims

Key fields include the claim, which outlines the core assertion, and the posted date, which provides temporal context. Additional features include the sci-digest with brief claim summaries, and the justification field, which offers reasoning for their validity. The dataset also includes visual elements through an image link and highlights claim complexities in the Issues column. The evidence field serves as the ground truth, validating the claims' accuracy.

The dataset consists of 1943 rows and 7 columns, offering a multidimensional resource that combines textual, chronological, evidential, and visual data. This robust framework supports the development of models capable of effectively detecting and explaining financial misinformation. Figure 2 illustrates the sector-wise distribution of claims.

4.2 Model Selection for Fine Tuning

We chose GPT-4o Mini for our fine-tuning based on thorough model evaluation metrics, providing a favorable trade-off between performance and computational efficiency. While GPT-4o Mini retains similar performance metrics (65/100 for both parameters) and dramatically lowers fine-tuning costs by about 60% and latency by 48%, the standard GPT-4o shows slightly better reasoning (67/100) and robustness (68/100) ratings. For our deployment scenario, where resource efficiency and model efficacy must be matched, this cost-performance optimization is essential.

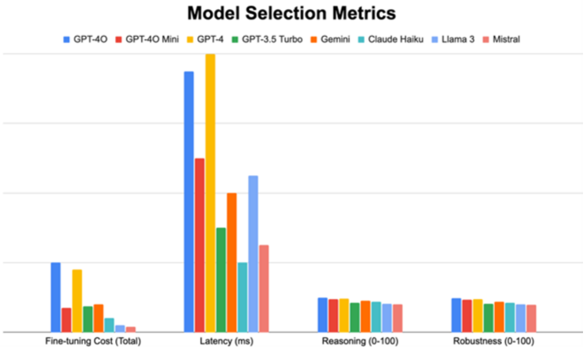

Figure 3: Model performance metrics across different experiments

Given the significant gains in computational efficiency and response times, the slight loss in reasoning and robustness capabilities (roughly 3% reduction) is a reasonable trade-off, making GPT-4o Mini the most practical option for our implementation needs.

5 Experimentation and Evaluation

The dataset was divided into training and validation sets using an 80-20 split, ensuring stratified sampling to preserve the class distribution across True, False, and NEI labels. This resulted in 1500 samples for training and 443 for validation. To ensure robust model evaluation, we implemented 5-fold cross-validation, providing insights into performance across different data splits.

Seven experiments were conducted to evaluate strategies for fine-tuning and prompt engineering, adapting a large language model (LLM) to the tasks of verifying financial claims and generating explanations (cf. Table 1). The task involved classifying claims and generating structured explanations aligned with an instruction prompt.

Experiment 1 used only prompt engineering without modifying the base model. While this approach achieved a high overall score (0.8348), task-specific metrics like F1 Micro (0.2247) and ROUGE1 (0.2225) were low, indicating limitations in aligning the LLM's reasoning with the problem domain.

Experiment 2 fine-tuned the base model using GPT-4o. This reduced the overall score to 0.5804 but significantly improved F1 Micro to 0.8706, suggesting enhanced claim categorization. However,

S.No	Experiment Specification	Overall Score	F1 Micro	Rouge 1	Rouge 2	Rouge L
1	GPT4o-mini with only prompt engineering	0.529	0.835	0.225	0.222	0.225
2	Fine tuned GPT4o-mini - 1st fine tuning attempt	0.580	0.871	0.290	0.113	0.179
3	Combined prompt engineering with fine tuned GPT4o-mini	0.603	0.879	0.326	0.221	0.257
4	Fine tuned GPT4o-mini with chaining prompt engineering	0.687	0.880	0.495	0.477	0.489
5	Fine tuned GPT4o-mini with more columns - 2nd attempt at fine tuning	0.692	0.879	0.505	0.409	0.428
6	Prompt engineering with updated fine tuned GPT4o-mini model	0.700	0.880	0.510	0.420	0.440
7	Proposed approach	0.763	0.903	0.623	0.440	0.460

Table 1: Evaluation results across different experimental settings

explanation generation required further refinement through prompting strategies.

Experiment 3 applied a single-layer prompting strategy with the fine-tuned model, yielding balanced improvements in ROUGE metrics (ROUGE1: 0.3267) and an overall score of 0.6033.

Experiment 4 introduced two-layer prompting, structuring intermediate reasoning steps to align better with task objectives. This approach improved ROUGE1 (0.4948) and ROUGE2 (0.4771), with an overall score of 0.6873.

Experiment 5 enhanced the fine-tuned model by incorporating synonym retrieval and lemmatization. This further improved ROUGE1 (0.5059) and stabilized the overall score at 0.6929.

Experiment 6 achieved the best task-specific performance by systematically improving the prompt template. This experiment recorded the highest overall score (0.6974) and ROUGE1 (0.5149), demonstrating the importance of refined prompt engineering.

Experiment 7 utilized a two-step framework. In the first step, multimodal attributes were added by extracting image content and URL summaries using AI tools, which were then used to retrain the model. In the second step, a different model reviewed and updated the explanations and labels from the first step. This approach achieved a high F1 score (0.90), highlighting the effectiveness of integrating multimodal data into the evaluation process.

These findings underscore the importance of harmonizing task-specific fine-tuning with iterative prompt design to achieve robust performance in both claim classification and explanation generation.

6 Conclusion

This paper introduces a novel two-step framework for detecting financial misinformation, effectively combining the strengths of fine-tuned Large Language Models (LLMs) with explainable AI principles. Our approach achieves state-of-the-art performance with an F1 score of 0.90, while ensuring transparency through detailed explanations of its decision-making process. The dual-layer verification system, which includes an LLM judge, significantly enhances the reliability of classifications and provides clear, actionable insights for financial stakeholders.

Our findings demonstrate that combining sophisticated prompt engineering with targeted fine-tuning yields superior performance compared to using either approach alone. Additionally, integrating multimodal attributes in the final experiment further improved the model's ability to accurately contextualize and verify financial claims.

7 Limitations

Despite the strong performance of our framework, several limitations should be acknowledged:

- Computational Resources: The two-step verification process increases computational overhead, which may impact real-time processing capabilities.

- Temporal Relevance: Financial markets are dynamic, requiring regular model updates to maintain accuracy with changing conditions and new forms of misinformation.

- Language Dependency: The current implementation focuses primarily on English-language content, limiting its global applicability.

- Cost Considerations: The use of GPT-4 based models, while effective, may pose cost barriers for smaller organizations or individual researchers.

These limitations present opportunities for future research, particularly in developing more efficient verification mechanisms and expanding the model's generalizability.

References

Perry Carpenter. 2023. Council post: Get the 411 on misinformation, disinformation and malinformation. *Forbes*.

S. Chen and T. Zhang. 2023. Fingpt: Open-source financial large language models for misinformation detection. In *Proceedings of the 2023 IEEE International Conference on Financial Technologies (FinTech)*, pages 1–10.

Wingyan Chung, Yinqiang Zhang, and Jia Pan. 2022. A theory-based deep-learning approach to detecting disinformation in financial social media. *Information Systems Frontiers*, pages 1–20.

Zhiwei Liu, Xin Zhang, Kailai Yang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Fmdllama: Financial misinformation detection based on large language models. *Preprint*, arXiv:2409.16452.

Tyler Moore, Jie Han, and Richard Clayton. 2012. The postmodern ponzi scheme: Empirical analysis of high-yield investment programs. In *Financial Cryptography and Data Security: 16th International Conference*, pages 41–56. Springer.

Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu. 2024. Fin-fact: A benchmark dataset for multimodal financial fact checking and explanation generation. *Preprint*, arXiv:2309.08793.

Z. Wang and Y. Liu. 2022. Crossfin: A cross-platform financial misinformation detection framework. In *Proceedings of the 2022 International Conference on Machine Learning and Data Engineering*, pages 40–50.

Z. Xie, X. Zhang, and Y. Wang. 2021. Fnfnet: Deep learning for fake news detection using convolutional neural networks. *IEEE Access*, 9:9634064.

H. Yang and C. Zhang. 2020. Finbert: Financial sentiment analysis with pre-trained language models. In *Proceedings of the 2020 IEEE/ACM International Conference on Advances in Financial Technology (FinTech)*, pages 1–7.

J. Yang and X. Liu. 2023. Dfdr: Deep financial disinformation recognition using multi-source market data. *International Journal of Computational Finance*, 15(3):250–263.

L. Zhang and M. Wang. 2023. Temporal-aware language models for financial misinformation detection. In *Proceedings of the 2023 IEEE International Conference on Financial Technology (FinTech)*, pages 112–120.

L. Zheng. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.