FinNLP-FNP-LLMFinLegal-2025 Shared Task: Financial Misinformation Detection Challenge Task

Zhiwei Liu[1,2] **Keyi Wang**[3,4] **Zhuo Bao**[5] **Xin Zhang**[1] **Jiping Dong**[6] **Kailai Yang**[1,2]

Mohsinul Kabir[1] **Polydoros Giannouris**[1] **Rui Xing**[7] **Seongchan Park**[8] **Jaehong Kim**[8]

Dong Li[9] **Qianqian Xie**[9] **Sophia Ananiadou**[1,2,10]

[1] University of Manchester [2] Artificial Intelligence Research Center [3] Columbia University
[4] Northwestern University [5] Internet Domain Name System Beijing Engineering Research Center Co
[6] University of Chinese Academy of Sciences [7] University of Melbourne
[8] Korea Advanced Institute of Science and Technology (KAIST) [9] The FinAI
[10] Archimedes/Athena RC

{zhiwei.liu,kailai.yang,sophia.ananiadou}@manchester.ac.uk
{xin.zhang-41,mdmohsinul.kabir,polydoros.giannouris}@postgrad.manchester.ac.uk
kw2914@columbia.edu baozhuo@zdns.cn dongjiping19@mails.ucas.ac.cn
ruixing@student.unimelb.edu.au {sc.park,luke.4.18}@kaist.ac.kr
{dong.li,qianqian.xie}@thefin.ai

Abstract

Despite the promise of large language models (LLMs) in finance, their capabilities for financial misinformation detection (FMD) remain largely unexplored. To evaluate the capabilities of LLMs in FMD task, we introduce the financial misinformation detection shared task featured at COLING FinNLP-FNP-LLMFinLegal-2024, FMD Challenge. This challenge aims to evaluate the ability of LLMs to verify financial misinformation while generating plausible explanations. In this paper, we provide an overview of this task and dataset, summarize participants' methods, and present their experimental evaluations, highlighting the effectiveness of LLMs in addressing the FMD task. To the best of our knowledge, the FMD Challenge is one of the first challenges for assessing LLMs in the field of FMD. Therefore, we provide detailed observations and draw conclusions for the future development of this field.

1 Introduction

The joint workshop of FinNLP, FNP, and LLMFinLegal aims to explore the intersection of Natural Language Processing (NLP), Machine Learning (ML), and Large Language Models (LLMs) within the financial and legal domains. In recent years, the FinNLP, FNP, and LLMFinLegal series have extensively investigated the intersection of FinTech, NLP, and LLMs. These efforts have systematically uncovered key challenges, provided strategic guidance for future research directions, and proposed a range of shared tasks within the financial domain, including sentence boundary detection (Azzi et al., 2019); (Au et al., 2020), learning semantic representations (Maarouf et al., 2020), semantic similarities (Kang et al., 2021; Kang and El Maarouf, 2022; Chen et al., 2023), and LLMs-based financial task (Xie et al., 2024).

In the financial sector, maintaining the accuracy of information is fundamental to ensuring market stability, supporting informed decision-making, managing risks effectively, fostering trust, and achieving regulatory compliance (Rangapur et al., 2023b). However, the widespread adoption of digital media has significantly exacerbated the dissemination of financial misinformation (Chung et al., 2023). Such misinformation, including biased news reports and deceptive investment schemes, poses considerable risks by influencing economic sentiment and distorting market prices (Kogan et al., 2020). Manual detection of financial misinformation is time-consuming and costly (Kamal

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 271–276
January 19–20, 2025.

et al., 2023), making automated detection a crucial research area. Additionally, ensuring the explainability of models in their decisions to identify misinformation enhances transparency, trust, and practical value for investors, regulators, and the financial community (Fritz-Morgenthal et al., 2022). The advent of LLMs in finance has introduced the transformative potential for analysis (Shah et al., 2022), prediction (Wu et al., 2023), and decision-making (Xie et al., 2023). However, few studies based on LLMs have focused on the critical field of financial misinformation detection.

To explore the ability of LLMs for financial misinformation detection, we propose the financial misinformation detection challenge shared task (FMD). This challenge includes one published dataset designed to address the financial misinformation detection challenge. We utilize the FMDID dataset (Liu et al., 2024), which is based on FinFact (Rangapur et al., 2023a). It is a comprehensive collection of financial claims categorized into various areas. Using this data, we design a prompt query template to adapt LLMs to identify financial claims and give explanations for their decision according to the related information.

This paper overviews the shared task and dataset in the FMD Challenge, summarizes participant methods, and evaluates their experiments to explore LLM's capabilities in financial misinformation detection. Our comprehensive evaluation highlights the strengths and limitations of current methodologies, showcasing the effectiveness of LLMs and the potential of domain-specific instruction tuning in the FMD task.

2 Task and Dataset

This task, derived from FMDID (Liu et al., 2024)'s FinFact (Rangapur et al., 2023a) part, a comprehensive collection of financial claims categorized into areas like Income, Finance, Economy, Budget, Taxes, and Debt. The claim label categorizes claims as "True", "False", and "NEI (Not Enough Information)". Table 1 presents the information in the dataset. The objective of this task is to evaluate the ability of LLM to verify financial misinformation while generating plausible explanations. The dataset includes 1952 training data and 1304 test data.

For instruction-tuning data, we use the following base prompt template example to support the training and evaluation of LLMs. Also, participants are encouraged to adjust the template to make full use of all information. *[task prompt]* denotes the instruction for the task (e.g. Please determine whether the claim is True, False, or Not Enough Information based on contextual information, and provide an appropriate explanation.). *[claim]* and *[context]* are the claim text and contextualization content from the raw data respectively. *[output1]* and *[output2]* are the outputs from LLM.

Feature	Notes
claim	the core assertion.
posted date	temporal information.
sci-digest	claim summaries
justification	justification offers insights into their accuracy to further contextualize claims.
image link	visual information.
issues	highlight complexities within claims.
evidence	supporting information, which serves as the ground truth of explanations

Table 1: The contents included in the dataset.

> **Task:** *[task prompt]*. **Claim:** *[claim]*. **Context:** *[context]*. **Prediction:** *[output1]*. **Explanation:** *[output2]*

This task adopts Micro-F1 for misinformation detection evaluation and ROUGE (1, 2, and L) (Lin, 2004) for explanation evaluation. The average of F1 and ROUGE -1 scores is applied as the final ranking metrics.

3 Participants and Automatic Evaluation

32 teams have registered for the FMD Challenge, out of which 8 teams have submitted their LLMs solution papers. We employ two baseline models from FMDLlama (Liu et al., 2024) and GPT-3.5-turbo[1]. FMDLlama is an open-sourced instruction following LLM for FMD task based on finetuning Llama3.1 with instruction data. GPT-3.5-turbo is one typical variant of OpenAI's products.

During the testing phase, we conducted the automatic evaluation using the Hugging Face platform[2]. We randomly selected 40% of the test dataset for the public evaluation phase, while the remaining 60% was designated as a private dataset. The score

[1]https://openai.com/

[2]https://huggingface.co/spaces/TheFinAI/FMD2025

Rank	Team name	overall score	micro-F1	rouge1	rouge2	rougeL
1	Dunamu ML	0.8294	0.8467	0.8121	0.7873	0.7969
2	GGbond	0.7924	0.7955	0.7892	0.7517	0.7663
3	1-800-SHARED-TASKS	0.7768	0.8283	0.7253	0.6763	0.6911
4	Drocks	0.7653	0.7877	0.7429	0.6983	0.7142
5	GMU-MU	0.6682	0.7575	0.5789	0.4956	0.5145
6	Ask Asper	0.6465	0.7824	0.5106	0.4025	0.4221
7	Team FMD LLM	0.5813	0.6448	0.5178	0.4428	0.4607
8	Capybara	0.5127	0.7221	0.3033	0.1014	0.174
Baseline	FMDLlama	0.5842	0.7182	0.4502	0.3464	0.3743
Baseline	ChatGPT (gpt-3.5-turbo)	0.4813	0.7012	0.2614	0.0994	0.1632

Table 2: Evaluation results on FMD challenge

on the public split was shown on the leaderboard in real-time. The score on the private split was shown after the deadline. The final rankings are based on the private split performance. Table 2 shows the final ranking and results.

4 Methods of Each Team

In this section, we provide a detailed overview of the LLMs-based solutions for each paper.

Dunamu ML employs data augmentation using a general-domain misinformation dataset, MOCHEG, to address data scarcity in the financial domain. They first collect claims and labels, generate evidence, and then construct few-shot examples on the augmented data based on sentence embedding similarity and perform supervised fine-tuning (SFT). Specifically, in the data augmentation process, GPT-4-0613 (Achiam et al., 2023) is first employed to generate evidence. For few-shot selection, OpenAI's text-embedding-3-large model is used to generate sentence embeddings, with cosine similarity serving as the similarity metric. Furthermore, the FAISS library (Douze et al., 2024) is utilized to perform the embedding similarity search. Finally, they fine-tune Llama-3.1-8B on the augmented dataset.

GGbond fine-tunes Llama 3.2-11B-Vision-Instruct (Dubey et al., 2024) using both text and image information. They first design specialized prompts to enable the Llama3.2-Vision model to choose the most relevant image and convert it into corresponding textual descriptions, including image description, contextual information, and relevant details. They subsequently apply LoRA to fine-tune the Llama-3.2-11B-Vision-Instruct model on the processed data.

1-800-SHARED-TASKS trains various LLMs through a sequential fine-tuning approach. They begin by fine-tuning five open-source LLMs (i.e. Qwen2.5 (Team, 2024), LLama3 8B (Dubey et al., 2024), Mistral 7B (Jiang et al., 2023), Phi3 medium 4K Instruct (Abdin et al., 2024), and Gemma-2 9B (Team et al., 2024)) exclusively for classification, then select the best-performing models for a second stage of fine-tuning for joint classification and explanation generation.

Drocks enhances GPT-4o-mini (Achiam et al., 2023) through instruction tuning and compares their results with various LLMs, including Vicuna-7b-v1.55 (Chiang et al., 2023), Mistral-7b-Instruct (Jiang et al., 2023), LLaMA2-chat-7b (Touvron et al., 2023), and LLaMA3.1-8b-Instruct (Dubey et al., 2024), ChatGPT (Achiam et al., 2023) and GPT-4o-mini (OpenAI).

GMU-MU fine-tunes Llama-3.1-8B (Dubey et al., 2024) directly using the datasets and they also compare with a few-shot prompt method. In the prompt method, they first ask the model to identify the main assertion or claim spans from both the claim and the associated context to generate a veracity label. The model subsequently provide a explanation for the predicted label while considering the claim and the associated context.

Ask Asper introduces a two-step framework utilizing GPT-4o-mini. They first fine-tune GPT-4o-mini on the financial datasets to classify claims and generate explanations. To enhance the reliability and accuracy of the initial stage, a second model serves as a verification layer, examining and refining the initial model's predictions and explanations.

Team FMD LLM fine-tunes Llama-3.2-1B-Instruct and explores the impact of two factors. The first is label prediction order. They compare whether classifying a financial claim

(True/False/NEI) before generating an explanation yields better performance than the reverse. Additionally, they also explore the potential benefits of leveraging auxiliary metadata, particularly the availability of the sci_digest field, which demonstrated a strong correlation with the labels.

Capybara combines retrieved evidence with a financial Chain-of-Thought prompt to enhance various LLMs. Specifically, they first apply one search engine (i.e. SerpAPI[3]) to retrieve the summarized information as supporting evidence. Subsequently, they introduce the Financial Chain of Thought (Financial CoT) from three dimensions: Alignment, Accuracy, and Generalization. This framework is designed to guide LLMs to focus more on reasoning during predictions, thereby enhancing their reasoning capabilities in the specific context of financial information.

5 Discussion

As shown in Table 2, the experimental results highlight the remarkable performance of various teams in the FMD task, especially for those that employed fine-tuning strategy with task-specific training data. Notably, *Dunamu ML* enhances the general-domain misinformation dataset with generated evidence and fine-tuned Llama-3.1-8B, achieving the best performance across all metrics. This highlights the importance of supplementing LLMs with additional structured knowledge to improve their task comprehension. Followed by *GGbond* and *1-800-SHARED-TASKS*, who make full use of both textual and visual information and one sequential fine-tuning approach respectively. From the results of *Drocks* and *GMU-MU*, it can be seen that directly fine-tuning LLMs with appropriately designed prompts can achieve relatively good overall performance. However, if the prompt design is inappropriate, the base model selection is not large, or the optimization strategy is unsuitable, the explanation generation capabilities may be highly sensitive and negatively affected (e.g. the rouge score of *GMU-MU*). This could also explain why the remaining teams achieved high scores in the classification task, but performed averagely in the explanation generation task. It is worth mentioning that *Capybara* replaced fine-tuning with evidence retrieval and the use of Financial CoT. Currently, it is indeed a challenge for LLMs to outperform fine-tuned models on specific tasks. Although it did not achieve a high score, it is worth exploring further, as it could help reduce the use of computational resources.

Overall, supplementing appropriate additional knowledge, utilizing multimodal information, or improving model size can enhance the performance of LLMs on specific tasks. Moreover, exploring alternatives to fine-tuning LLMs is also worth further consideration.

6 Conclusion

In this paper, the FMD Challenge has demonstrated the efficacy and potential of LLMs in the domain of financial misinformation detection. Our challenge, along with the resources provided, has significantly contributed to advancing this field. Participants utilized these resources to develop effective strategies and models, which led to improved performance. The experimental results highlight the considerable value of LLMs-based approaches. The overall trend indicates that performance improves with increasing model size and advancements in fine-tuning and prompt engineering. These findings offer valuable insights for future research in FMD task using LLMs. The success of this challenge underscores the importance and impact of collaborative efforts in pushing the boundaries of AI applications in finance.

Acknowledgments

We would like to thank all the anonymous reviewers and area chairs for their comments. This work is supported by the project JPNP20006 from New Energy and Industrial Technology Development Organization (NEDO) and the Manchester-Melbourne-Toronto Research Funding.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Willy Au, Bianca Chong, Abderrahim Ait Azzi, and Dialekti Valsamou-Stanislawski. 2020. FinSBD-2020:

[3]https://serpapi.com/

The 2nd shared task on sentence boundary detection in unstructured text in the financial domain. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 47–54, Kyoto, Japan. -.

Abderrahim Ait Azzi, Houda Bouamor, and Sira Ferradans. 2019. The FinSBD-2019 shared task: Sentence boundary detection in PDF noisy text in the financial domain. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pages 74–80, Macao, China.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Yohei Seki, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023. Multi-lingual ESG impact type identification. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, pages 46–50, Bali, Indonesia. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, march 2023. *URL https://lmsys. org/blog/2023-03-30-vicuna*, 3(5).

Wingyan Chung, Yinqiang Zhang, and Jia Pan. 2023. A theory-based deep-learning approach to detecting disinformation in financial social media. *Information Systems Frontiers*, 25(2):473–492.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Sebastian Fritz-Morgenthal, Bernhard Hein, and Jochen Papenbrock. 2022. Financial risk management and explainable, trustworthy, responsible ai. *Frontiers in artificial intelligence*, 5:779799.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Ashraf Kamal, Padmapriya Mohankumar, and Vishal Kumar Singh. 2023. Financial misinformation detection via roberta and multi-channel networks. In *International Conference on Pattern Recognition and Machine Intelligence*, pages 646–653. Springer.

Juyeon Kang and Ismail El Maarouf. 2022. FinSim4-ESG shared task: Learning semantic similarities for the financial domain. extended edition to ESG insights. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 211–217, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Juyeon Kang, Ismail El Maarouf, Sandra Bellato, and Mei Gan. 2021. FinSim-3: The 3rd shared task on learning semantic similarities for the financial domain. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, pages 31–35, Online. -.

Shimon Kogan, Tobias J Moskowitz, and Marina Niessner. 2020. *Fake news in financial markets*. SSRN.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Ismail El Maarouf, Youness Mansar, Virginie Mouilleron, and Dialekti Valsamou-Stanislawski. 2020. The FinSim 2020 shared task: Learning semantic representations for the financial domain. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 81–86, Kyoto, Japan. -.

Gpt OpenAI. 4o mini: advancing cost-efficient intelligence, 2024. *URL: https://openai. com/index/gpt-4o-mini-advancing-cost-efficient-intelligence*.

Aman Rangapur, Haoran Wang, and Kai Shu. 2023a. Fin-fact: A benchmark dataset for multimodal financial fact checking and explanation generation. *Preprint*, arXiv:2309.08793.

Aman Rangapur, Haoran Wang, and Kai Shu. 2023b. Investigating online financial misinformation and its consequences: A computational perspective. *arXiv preprint arXiv:2309.12363*.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *arXiv preprint arXiv:2211.00083*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.

Qianqian Xie, Jimin Huang, Dong Li, Zhengyu Chen, Ruoyu Xiang, Mengxi Xiao, Yangyang Yu, Vijayasai Somasundaram, Kailai Yang, Chenhan Yuan, et al. 2024. Finnlp-agentscen-2024 shared task: Financial challenges in large language models-finllms. In *Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning*, pages 119–126.