FMDLlama: Financial Misinformation Detection based on Large Language Models

Zhiwei Liu 1 Xin Zhang 1 Kailai Yang 1 Qianqian Xie 2 Jimin Huang 2 Sophia Ananiadou 1,3,4

¹ The University of Manchester ² The Fin AI

³ Artificial Intelligence Research Center ⁴ Archimedes/Athena RC

{zhiwei.liu,kailai.yang,sophia.ananiadou}@manchester.ac.uk

xin.zhang-41@postgrad.manchester.ac.uk, {qianqian.xie,jimin.huang}@thefin.ai

Abstract

The emergence of social media has made the spread of misinformation easier. In the financial domain, the accuracy of information is crucial for various aspects of financial market, which has made financial misinformation detection (FMD) an urgent problem that needs to be addressed. Large language models (LLMs) have demonstrated outstanding performance in various fields. However, current studies mostly rely on traditional methods and have not explored the application of LLMs in the field of FMD. The main reason is the lack of FMD instruction tuning datasets and evaluation benchmarks. In this paper, we propose FMDLlama, the first open-sourced instructionfollowing LLMs for FMD task based on finetuning Llama3.1 with instruction data, the first multi-task FMD instruction dataset (FMDID) to support LLM instruction tuning, and a comprehensive FMD evaluation benchmark (FMD-B) with classification and explanation generation tasks to test the FMD ability of LLMs. We compare our models with a variety of LLMs on FMD-B, where our model outperforms all other open-sourced LLMs as well as ChatGPT.

1 Introduction

In the financial sector, the accuracy of information is crucial for the integrity of decisions, market operation, risk management, compliance, and trust establishment (Rangapur et al., 2023b). However, the proliferation of digital media has escalated the spread of financial misinformation (Chung et al., 2023). Such misinformation, including deceptive investment propositions and biased news articles, can manipulate market prices and influence economic sentiment, presenting substantial risks (Kogan et al., 2020). Furthermore, manually checking financial misinformation consumes a large amount of time and manpower (Kamal et al., 2023). Therefore, the automatic identification of financial misinformation is an urgent priority for the normal

operation of financial activities, yet there is currently limited exploration by researchers in this area.

Recently, large Language Models (LLMs) with large parameters have been explored as a new approach to address various issues in the financial domain (e.g. analysis (Shah et al., 2022), prediction (Wu et al., 2023), and decision-making (Xie et al., 2023)), yielding promising results. However, most studies concentrate on applying traditional deep learning methods such as CNNs, LSTMs, or pretrained language models (PLMs) with fewer parameters like BERT or RoBERTa to detect false information in the financial domain (Kamal et al., 2023; Chung et al., 2023; Mohankumar et al., 2023). Rangapur et al. (2023a) evaluate a few LLMs on the FMD task. There are currently no LLMs specifically designed for detecting financial misinformation. The main reason is the lack of data available for instruction-tuning LLMs.

To address the above issues, we construct the first instruction-tuning datasets for financial misinformation detection (FMDID) to support LLMs fine-tuning, including classification and explanation tasks. We subsequently propose FMDLLMs, the first open-sourced financial misinformation detection LLMs (FMDLlama) based on FMDID. To evaluate the financial misinformation verification ability of LLMs, we also build a benchmark for the detection of financial misinformation (FMD-B). The results on FMD-B show that FMDLLMs achieve state-of-the-art (SOTA) performance among other open-sourced LLMs, as well as the closed-sourced ChatGPT.

Our main contributions are as follows:

- (1) We construct the FMDID, the first multi-task financial misinformation instruction-tuning dataset.
- (2) We develop FMDLLMs, the first opensourced financial misinformation detection LLMs that are specialized for diverse financial misinformation detection tasks.

(3) We build FMD-B, the first benchmark to evaluate the verification ability of financial misinformation of LLMs. The results on FMD-B demonstrate that our model overtakes other open-sourced LLMs and ChatGPT.

2 Related work

2.1 Financial Misinformation Detection

There are few studies detecting financial misinformation. Most of them are based on traditional deep learning methods or PLMs. (Kamal et al., 2023) introduce a framework for FMD task based on RoBERTa and multi-channel networks (CNNs, BiGRU, and attention layers). (Chung et al., 2023) apply multiple LSTMs to learn dynamic and hidden patterns to support financial disinformation detection. (Mohankumar et al., 2023) adopt two cross-joint networks to build contextual sequential representation, which is produced by the combination of context-aware linguistic and financial embeddings, to detect fake news. Rangapur et al. (2023a) propose one dataset for financial fact checking and explanation generation, and evaluate the ability of several LLMs (e.g. GPT-4, Claude3, Mistral) on this dataset. However, there is currently no open-sourced LLM specifically designed for the detection of financial misinformation.

2.2 Open Sourced Large Language Models

Significant research efforts have focused on creating open-sourced LLMs as alternatives to the closed-sourced models (e.g. ChatGPT, GPT-4), which aims to facilitate more accessible research into enhancing and applying LLMs. Well-known series of open-sourced, general-purpose language models include LLaMA series (e.g. llama2, llama3, llama3.1) (Touvron et al., 2023), Vicuna-7B-v1.5¹, Gemma (Team et al., 2024), Mistral (Jiang et al., 2023). There are also many open-sourced LLMs for specific domains, including FinMA (Xie et al., 2023) for finance, MentalLLaMA (Yang et al., 2023) for mental health, ExTES-LLaMA (Zheng et al., 2023) for emotional support chatbots, and EmoLLMs (Liu et al., 2024b) for sentiment analysis and ConspEmoLLM (Liu et al., 2024a) for conspiracy detection. In this work, we extend the inventory of domain-specific LLMs, by developing the first open-sourced LLM for multitask financial misinformation detection.

3 Methods

3.1 Task formalization

We approach financial misinformation detection as a generative task, applying a generative model as a foundation. This generative model is an autoregressive language model $P_{\phi}(y|x)$, parameterized using pre-trained weights ϕ . It has the ability to simultaneously handle multiple financial misinformation detection tasks, i.e., misinformation detection, and explanation generation. Each task (t) is represented as a set of context-target pairs: $D_t = (q_i^t, r_i^t)_{i=1,2,\dots,N_t}$, where the context q is a token sequence containing the task description, input text, and query, and r is a further token sequence containing the answer to the query. The model is optimized based on the merged dataset, which combines all task datasets, with the aim of maximizing the objective of conditional language modeling to improve prediction and generation performance.

3.2 Construction of instruction tuning dataset

Data		Raw		Instruction				
	Train	Val	Test	Train	Val	Test		
FinFact	1562	391	1304	1562	391	1304		
FinGuard	2900	600	1500	2900	600	1500		

Table 1: Dataset statistics. *Raw* denotes the raw data from FinFact and FinGuard. *Instruction* denotes the converted instruction data based on *Raw*.

3.2.1 Raw data

We build our instruction tuning dataset using two existing datasets.

FinFact FinFact (Rangapur et al., 2023a) is 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)'. It is meticulously crafted to reflect the complexity of financial narratives, including contextual details, supporting evidence links, and visual elements like image links and captions for each claim. A distinctive feature of this dataset is that it provides explanations for why each claim is deemed true or false, enriching the dataset's utility for training LLMs in not just detecting misinformation but also articulating reasoned explanations for their assessments.

FinGuard Financial Truth Guard dataset² is

¹https://huggingface.co/lmsys/vicuna-13b-v1.5

 $^{^2} https://github.com/carlos-gmartin/Financial-Truth-Guard \\$

Task	Instruction Template
FinFact FinGuard	Task: Please determine whether the claim is 0. False, 1. True, or 2. Not Enough Information (NEI)
	based on contextual information, and provide an appropriate explanation.
	The answer needs to use the following format:
	Prediction: [0. False, 1. True, or 2. NEI]
	Explanation: [Explain why the above prediction was made]
	Claim: [raw claim]. Claim summaries: [raw summaries]. Contextual information: [raw contextual]
	Task: Please determine whether the text is 0. Fake or 1. True. Answer directly without explanations.
	Text: [input text]

Table 2: Instructions used for each task.

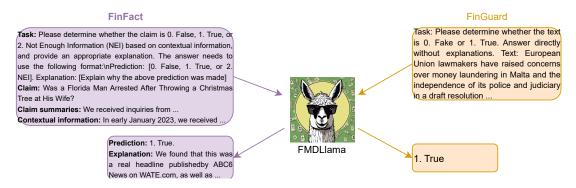


Figure 1: An overview of multi-task instruction tuning of FMDLlama

used for analyzing news articles and predicting whether they contain misleading or fake information related to the financial markets.

3.2.2 Construction of the FMD instruction tuning dataset (FMDID) and FMD benchmark (FMD-B)

We use the raw datasets as the basis to build the instruction dataset. For FinFact, We split the original data into train, validation, and test sets and remove the data without evidence (i.e. explanation). For FinGuard, we separately extracted 2500 data points from both real and fake data and divided them into the train, validation, and test sets. The dataset statistics are presented in Table 1. We construct instruction-tuning data for each task based on the template in Table 2. For FinFact, [raw claim], [raw summaries], and [raw contextual] are from the raw data. The format of LLMs' response will be Prediction: [Lable]. Explanation: [Explanations]. [Lable] is one of [0. False, 1. True, or 2. NEI]. [Explanations] is the reason why the LLM makes the [Lable] decision. For FinFact, [Explanations] is from the raw data. Figure 1 presents examples used to fine-tune the LLM.

After constructing the instruction data. We collect the train and validation data as instruction-tuning data (FMDID) and test data as the Financial Misinformation Detection benchmark (FMD-B),

which are used to fine-tune the LLMs and evaluate the ability of LLMs in the Financial Misinformation Detection domain respectively.

3.3 FMDLLMs

We built FMDLlama2 and FMDLlama3 by fine-tuning LLaMA2-chat-7b (Touvron et al., 2023) and Llama-3.1-8B-Instruct³ using the FMDID dataset. The models are trained based on the AdamW optimizer (Loshchilov and Hutter, 2017) for three epochs, using DeepSpeed (Rasley et al., 2020) to reduce memory usage. We set the batch size to 128. The initial learning rate is set to 1e-6 with a warm-up ratio of 5%. All models are trained on two Nvidia Tesla A100 GPUs, each with 80GB of memory. Figure 1 provides an overview of multitask instruction tuning of FMDLlama for diverse financial misinformation detection tasks.

4 Experiments

4.1 Baseline models

PLMs: Financial misinformation detection is typically regarded as a classification task. For our baseline models, we select commonly used PLMs, which can only be fine-tuned for individual tasks, i.e., the general language BERT (Devlin et al.,

³https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct

-	FinFact							FinGuard				
	Classification				Explanation				Classification			
	ACC	PRE	REC	F1	R1	R2	RL	BERTscore	ACC	PRE	REC	F1
BERT	0.6221	0.5550	0.5153	0.4836	-	-	-	-	0.9845	0.9845	0.9845	0.9845
RoBERTa	0.6822	0.6373	0.5823	0.5661	-	-	-	-	0.9961	0.9962	0.9961	0.9961
gemma-2b	0.0782	0.4915	0.0782	0.0782	0.0295	0.0077	0.0208	0.0539	0.1047	0.5086	0.1047	0.1047
vicuna-7b	0.3758	0.7292	0.3758	0.3758	0.2712	0.1012	0.1677	0.5429	0.4167	0.4759	0.4167	0.4167
vicuna-13b	0.2991	0.5132	0.2991	0.2991	0.2593	0.0884	0.1568	0.5310	0.3707	0.5415	0.3707	0.3707
mistral-7b	0.6097	0.7214	0.6097	0.6097	0.2724	0.0859	0.1574	0.5492	0.5887	0.5903	0.5887	0.5887
llama2-7b	0.3198	0.3720	0.3198	0.3198	0.1557	0.0495	0.0950	0.3443	0.3100	0.5598	0.3100	0.3100
llama2-13b	0.1933	0.5139	0.1933	0.1933	0.0645	0.0181	0.0399	0.1538	0.3053	0.5045	0.3053	0.3053
llama3.1-8b	0.6449	0.6494	0.6449	0.6449	0.2111	0.0823	0.1394	0.5449	0.5600	0.6513	0.5600	0.5600
ChatGPT	0.7270	0.7021	0.7270	0.7270	0.2639	0.1005	0.1645	0.5642	0.6800	0.7392	0.6800	0.6800
FMDllama2	0.6986	0.6886	0.6986	0.6986	0.2443	0.1670	0.2011	0.5214	0.9433	0.9654	0.9433	0.9433
FMDLlama3	0.7362	0.7211	0.7362	0.7362	0.4524	0.3498	0.3773	0.6756	0.9947	0.9947	0.9947	0.9947

Table 3: Results on FMD-B. R1, R2, RL denote ROUGE (1, 2, and L) respectively.

2018) and RoBERTa (Liu et al., 2019). We treat FinFact as a 3-way classification task, and FinGuard as a binary classification task, using cross-entropy loss for training.

LLMs: LLMs have been proven to be capable of solving numerous tasks. We apply zeroshot prompting on the instruction dataset to the following open-sourced LLMs: LLaMA2-chat-(7b,13b) (Touvron et al., 2023), LLaMA3.1-8b-Instruct⁴, Gemma-instruct-(2b, 7b) (Team et al., 2024), Mistral-7b-Instruct (Jiang et al., 2023) and Vicuna-(7b,13b)-v1.5⁵. We also utilize zero-shot prompting with the proprietary LLM ChatGPT.

4.2 Evaluation methods

We uses metrics such as Accuracy, Precision, Recall, Micro-F1 for misinformation detection (classification) evaluation and ROUGE (1, 2, and L) (Lin, 2004), BERTScore (Zhang et al., 2019) for explanation evaluation.

4.3 Results

Table 3 presents the results on FMD-B. From the table, we can see FMDLlama3 achieve SOTA results among all other open-sourced LLMs as well as the close-source ChatGPT. Although BERT and RoBERTa were fine-tuned on each classification task separately and have similar results with FMDLlama3 on simple FinGuard dataset, their performance is lower than FMDLlama3 on the complex dataset FinFact. A possible reason is that it is challenging for the PLMs with less parameters to understand long and complex textual con-

tent. For LLMs without fine-tuning, Mistral-7b, Llama3.1-8b, and ChatGPT perform well. This is because most of the data points in FMD-B are long texts. Mistral-7b, Llama3.1-8b and ChatGPT allow longer input lengths, and have a better understanding of corresponding long texts. By comparing the results of FMDLlama2 with llama2-7b and FMDLlama3 with llama3.1-8b, we can recognize the effectiveness of instruction-tuning strategies. Instruction-tuning strategies in specific domains allow LLMs to focus more on those domains, leading to better performance. Overall, the results from the table indicate that in the field of financial misinformation detection, our open-sourced model FMDLLaMa3 with 8b parameters has surpassed the closed-sourced ChatGPT with 170b parameters.

5 Conclusion

In this paper, we propose FMDLlama, the first LLM for financial misinformation detection (FMD). We also construct a multi-task FMD instruction dataset (FMDID) and a FMD evaluation benchmark (FMD-B). We conduct a comprehensive analysis of the performance of FMDLlama, as well as a variety of LLMs on the FMD-B benchmark. The results indicate that FMDLlama performs exceptionally well in FMD tasks, achieving SOTA compared to the other open-sourced LLMs as well as ChatGPT.

In the future, we aim to augment the FMDID and FMD-B datasets with further FMD datasets, including data from multiple platforms, sources, domains and languages, which can help further improve the FMDLlama and evaluate the FMD ability of LLMs more comprehensively.

⁴https://www.llama.com/

⁵https://huggingface.co/lmsys/vicuna-13b-v1.5

6 Limitations

The potential limitations of our work may be summarized as follows:

- (1) Due to restricted computational resources, we only carried out instruction-tuning/evaluation of financial misinformation detection tasks using 7b/13b LLMs. As such, we have not considered the impact of using larger models on the FMD tasks.
- (2) Due to the limited availability of publicly accessible datasets on financial misinformation, we constructed instruction-tuning datasets and benchmarks for financial misinformation detection based solely on two datasets, two kinds of tasks (i.e. classification, explanation generation).

References

- 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.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- 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.
- 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.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Zhiwei Liu, Boyang Liu, Paul Thompson, Kailai Yang, Raghav Jain, and Sophia Ananiadou. 2024a. Conspemollm: Conspiracy theory detection using an emotion-based large language model. *arXiv preprint arXiv:2403.06765*.

- Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024b. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487–5496
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Padmapriya Mohankumar, Ashraf Kamal, Vishal Kumar Singh, and Amrish Satish. 2023. Financial fake news detection via context-aware embedding and sequential representation using cross-joint networks. In 2023 15th International Conference on COMmunication Systems & NETworkS (COMSNETS), pages 780–784. IEEE.
- 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.
- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.
- 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.
- 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*.
- Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, and Sophia Ananiadou. 2023. Mentalllama: Interpretable mental health analysis on social media with large language models. *arXiv preprint arXiv:2309.13567*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*.