# Capybara at the Financial Misinformation Detection Challenge Task: Chain-of-Thought Enhanced Financial Misinformation Detection

**Yupeng Cao, Haohang Li, Yangyang Yu, Shashidhar Reddy Javaji**
Stevens Institute of Technology
Hoboken, NJ
{ycao33,hli113,yyu44,sjavaji}@stevens.edu

## Abstract

Financial misinformation poses a significant threat to investment decisions and market stability. Recently, the application of Large Language Models (LLMs) for detecting financial misinformation has gained considerable attention within the natural language processing (NLP) community. The Financial Misinformation Detection (FMD) challenge @ Coling 2025 serves as a valuable platform for collaboration and innovation. This paper presents our solution to FMD challenge. Our approach involves using search engines to retrieve the summarized high-quality information as supporting evidence and designing a financial domain-specific chain-of-thought to enhance the reasoning capabilities of LLMs. We evaluated our method on both commercial closed-source LLMs (GPT-family) and open-source models (Llama-3.1-8B and QWen). The experimental results domonstrate that the proposed method improves veracity prediction performance. However, the quality of the generated explanations remains relatively poor. In the paper, we present the experimental findings and provides an in depth analysis of these results.

## 1 Introduction

The proliferation of misinformation in the financial sector significantly impacts investor decision-making and market stability (Kogan et al., 2020; Liu and Moss, 2022). Manually verifying such financial misinformation demands substantial time and effort. Consequently, the development of automated tools for detecting financial misinformation has become a critical area of research in FinTech.

Previously, most frameworks for financial misinformation detection (FMD) relied on conventional deep learning approaches. For instance, (Kamal et al., 2023) developed a framework using RoBERTa combined with a multi-channel network (CNN, BiGRU, and an attention layer) specifically for FMD task, while (Chung et al., 2023) utilized multiple LSTMs to identify dynamic and covert patterns aiding in the detection process. Recently, with the advent of large language models (LLMs), in response to the complexity of the financial context and the professionalism of financial information, Fin-Fact (Rangapur et al., 2023) proposed a multimodal financial misinformation detection and interpretation generation dataset, and evaluated the capabilities of multiple popular LLMs on this dataset. Furthermore, FMDllama (Liu et al., 2024) has pioneered the use of open-source LLMs for identifying fraudulent financial information, setting a new benchmark in the field.

Despite these developments, the effectiveness of LLMs in FMD task warrants further exploration. The Financial Misinformation Detection Challenge @ COLING 2025, as introduced by FMDllama (Liu et al., 2024), aims to explore the capabilities of LLMs in enhancing the accuracy of financial misinformation detection. This paper describes our technical solution for FMD Challenge.

The core idea of our solution is to involves enhancing the "justification" component of the dataset by retrieving summarized high-quality information from online as the extra evidence using search engines and designing a financial domain-specific Chain-of-Thought Prompt to guide LLM reasoning and explanation generation. We conducted experiments on both commercial closed-source models and open-source models. Extensive evaluations on the FMD tasks yielded significant findings: (1) the proposed Financial Chain-of-Thought Prompt method effectively improves the pipeline's prediction results; and (2) despite this, the overall performance remains average. Furthermore, the quality of the generated explanations is significantly inferior to that of the baseline method, which has undergone fine-tuning. This underscores the necessity of fine-tuning the model using high-quality data. A more detailed analysis of the results is provided in Section 4.



*Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal*, pages 321–325
January 19–20, 2025. 

## 2 Shared Task Description

### 2.1 Problem Definition

The challenge focuses on developing advanced language models capable of detecting financial misinformation while providing explanatory justifications for their decisions. This dual objective—detection and explanation—represents a significant advancement over traditional binary classification approaches in financial text analysis. The task requires processing financial claims across diverse domains including income, finance, economics, budget, taxes, and debt. For each claim $c$, the model $M$ will take the query $q$ which includes claim $c$, justification $j$ and task description prompt $d$, and then model must make a three-way classification $y \in$ ['0. False', '1. True', or '2. Not Enough Information (NEI)'] and generate a coherent explanation $e$ supporting its decision. This explanation requirement adds a crucial layer of transparency and interpretability to the model's decision-making process, making it particularly valuable for real-world financial applications.

### 2.2 Challenge Dataset

The challenge utilizes the Fin-Fact (Rangapur et al., 2023) dataset, which provides rich contextual information for each financial claim, including temporal metadata, claim summaries, justifications, and supporting evidence. Participants are required to develop models that can effectively leverage this multi-faceted information to make accurate predictions while generating explanations that are both factual and well-reasoned. The challenge organizer also constructs the "instruction-following" version for fine-tuning usage. The datasets content can be found in following URL[1][2].

Performance evaluation employs a comprehensive metric framework combining classification accuracy measures (Accuracy, Precision, Recall, Micro-F1) with text generation quality metrics (ROUGE-1/2/L (Lin, 2004) and BERTScore (Zhang et al., 2019)). The final ranking is determined by averaging the F1 and ROUGE-1 scores, ensuring balanced assessment of both classification performance and explanation quality.

[1] https://huggingface.co/datasets/lzw1008/COLING25-FMD/tree/main/Training

[2] https://huggingface.co/datasets/lzw1008/COLING25-FMD

## 3 Methodology

In this section, we outlines the proposed pipeline for financial misinformation detection. We integrate the retrieved summarized high-quality information with the original justification as whole support information and utilize a Chain-of-Thought Prompt to enhance the prediction process (See in figure 1).

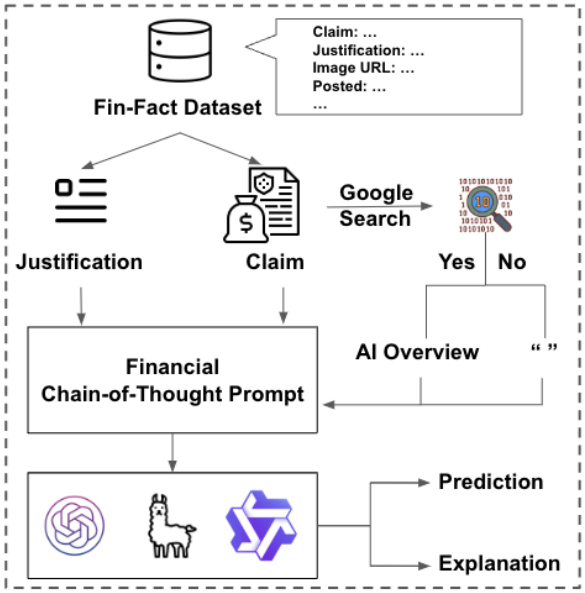


Figure 1: Schematic of proposed FMD pipeline.

### 3.1 Summarized High-Quality Evidence Retrieval

Previous research on fakenews detection and claim verification has shown that leveraging external verified knowledge sources, such as Wikipedia/Wikidata, can effectively authenticate information (Zhang and Ghorbani, 2020; Thorne et al., 2018; Aly et al., 2021). Recently, for more real-world wild claims, it becomes necessary to search the broader information from online for verification process (Schlichtkrull et al., 2023; Chen et al., 2023; Yue et al., 2024). This process, however, involves additional computational overhead for searching and post-processing the retrieved content. Financial-related claims pose unique challenges, as they often lack readily available information online due to their specialized domain knowledge and niche audience. Moreover, excessive information can introduce noise, potentially undermining prediction accuracy. Retrieving valid, high-quality information is therefore a challenge.

As search technology has evolved, Google Search Engine provides "AI Overview" results,

which are summaries automatic generated by search engine that combine data from various online sources and summarize them into concise information as output, aiming to efficiently answer queries. We utilize 'SerpAPI'[3] to search for each claim and concatenate the search results containing "AI Overview" with "Justification". If "AI Overview" is None, we keep the original content of "Justification". The search statistics are as follows:

| Dataset | No. of AI_Overview | No. of data |
|---|---|---|
| Practice Set | 31 | 600 |
| Train Set | 75 | 1953 |
| Test Set | 43 | 1304 |

Table 1: Number of results for 'AI Overview' compared to total number of data.

### 3.2 Chain-of-Thought for Financial Misinformation Detection

Chain-of-Thought prompting has demonstrated advantages across various reasoning tasks (Wei et al., 2022; Lyu et al., 2023). Inspired by that, We propose financial Chain-of-Thought (Financial CoT) from the following dimensions, tailored to the specific context of financial information, to guide large language models in focusing their reasoning during the prediction process, aiming to enhance their reasoning capabilities:

1. **Alignment:** Evaluate whether the claim content aligns in meaning with the provided evidence on the financial topic.
2. **Accuracy:** Check for accurate quantitative and qualitative representation of financial data, trends, or performance metrics mentioned in the claim.
3. **Generalization:** Identify any overgeneralization or oversimplification of financial trends, potentially misrepresenting unique cases as broader patterns.

The designed Financial Chain-of-Thought not only aids the LLMs in systematically dissecting and assessing factual content but also aligns their reasoning process with structured, human-like analytical methods. We combine the input claim, justification, and financial CoT to construct the input query, which is then fed into the LLM to simultaneously generate predictions and corresponding explanations. The whole Prompt is shown in below:

[3] https://serpapi.com/

**Financial CoT Prompt.**

**System Message:** You are a Fact Checker and You need to focus on the financial sector. Given a claim, assess the factual accuracy of the claim based on the evidence and generate the explaination. Please follow the steps below to think about making a prediction and provide an explanation for your prediction:
1. Alignment: Evaluate whether the claim content aligns in meaning with the provided evidence on the financial topic (e.g., stock performance, economic indicators).
2. Accuracy: Check for accurate quantitative and qualitative representation of financial data, trends, or performance metrics mentioned in the claim.
3. Generalization: Identify any overgeneralization or oversimplification of financial trends, potentially misrepresenting unique cases as broader patterns.

**User Message:** I will give you one claim and relevant evidence. Your task is to verify the factual authenticity of the claim based on the evidence provided. Make a final prediction from: 'True', 'False' or 'Not Enough Info' and provide a detailed explanation. Please provide the final output in JSON format containing the following two keys: prediction and explanation.

## 4 Experiment and Discussion

### 4.1 Experiment Setup

In this study, we employed closed-source models from the GPT family[4] and open-source models, including LLama3.1-8B-Instruct[5] and QWen2-7B-Instruct (Yang et al., 2024), as the backbone LLMs. The open-source models were run on a single NVIDIA RTX-A6000 GPU with 48GB DRAM. Additionally, we conducted a Zero-Shot Prompt (see in Appedix A.1) experiment for comparison. To ensure experimental reproducibility, the temperature was set to 0. The output length was uniformly set to 512 to generate valid explanations.

During the practice stage, we split the training set into a training portion and a validation portion

[4] https://openai.com/api/
[5] https://ai.meta.com/blog/meta-llama-3-1/

| Model | Zero-Shot Prompt | | | | | Financial CoT Prompt | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | Precision | Recall | F1 | Rouge-1 | Accuracy | Precision | Recall | F1 | Rouge-1 |
| Llama-3.1-8B-Instruct | 0.6449 | 0.6494 | 0.6449 | 0.6449 | 0.2111 | 0.7146 | 0.7405 | 0.6019 | 0.5541 | 0.1909 |
| QWen2-7B-Instruct | 0.6937 | 0.7940 | 0.5833 | 0.5201 | 0.1536 | 0.7028 | 0.8012 | 0.5888 | 0.5276 | 0.1662 |
| GPT-4o-mini | 0.7005 | 0.6856 | 0.6127 | 0.6241 | 0.3199 | 0.7175 | 0.6990 | 0.6447 | 0.6462 | 0.2971 |
| GPT-4o | 0.7342 | 0.7143 | 0.6538 | 0.6467 | **0.3341** | 0.7278 | 0.7253 | 0.7278 | **0.7221** | 0.3033 |
| GPT-4 | 0.7086 | 0.7086 | 0.7086 | 0.6680 | 0.3287 | 0.7131 | 0.7102 | 0.7131 | 0.6723 | 0.3097 |

Table 2: Overall Late Submission Results.

in an 90:10 ratio for performance evaluation. The models were subsequently tested and compared using the provided testing datasets.

### 4.2 FMD Challenge Results

We evaluated the performance of different models under zero-shot settings and with the Financial Chain-of-Thought (CoT) approach on a sampled validation set. Based on the evaluation results, we selected GPT-4o with the Financial CoT to conduct the final experiments on the competition test set and submitted the results. The Table 3 is leaderboard result: The evaluation results revealed that

| Overall Score | Micro F1 | Rouge 1 | Rouge 2 | Rouge L |
|---|---|---|---|---|
| 0.5127 | 0.7221 | 0.3033 | 0.1014 | 0.174 |

Table 3: The score of submitted results.

the Rouge scores were suboptimal, which negatively impacted the overall score. Consequently, we conducted additional tests after the challenge results were released to further evaluate our method.

### 4.3 Late Submission Results

The Table 2 compares the performance of various models under Zero-Shot Prompt and Financial CoT Prompt. We can find that Financial CoT prompt led to noticeable improvements across most metrics compared to the Zero-Shot Prompt setting. In detail, GPT-4o achieved the highest Recall (0.7278) and F1 Score (0.7221), demonstrating the robustness and effectiveness of the CoT Prompt approach. Similarly, GPT-4 showed robust performance with an F1 Score of 0.6723, indicating that CoT contributes positively to explanation quality. Furthermore, the closed-source models consistently outperformed the 7B/8B open-source models, indicating that models with larger parameter counts exhibit stronger reasoning performance. This observation aligns with the scaling-law trend, which suggests that increasing model size improves overall inference capabilities.

### 4.4 Analysis

Although the results demonstrate that the Financial CoT prompt significantly enhances model performance, the overall performance and leaderboard ranking remain suboptimal. Therefore, we conducted a more in-depth analysis. First, for the 7B/8B-level open-source models used in the experiment, the results under the zero-shot and Financial CoT settings were comparable, with the Llama-3.1-8B-Instruct model even performing better in the zero-shot setting. This indicates that the Financial CoT prompt is less effective when the inference capability of a smaller model is limited and may even disrupt the model's original reasoning process. Second, the overall Rouge scores were particularly unsatisfactory. The explanations generated with the Financial CoT prompt were worse than those produced directly by the model under the zero-shot setting, highlighting a significant gap between the generated explanations and human-like explanations. Compared with the baseline results (Liu et al., 2024), this suggests that additional fine-tuning steps may be necessary to improve performance. In addition, we observed during the experiment that the open-source models occasionally failed to generate responses effectively, requiring repeated attempts to produce a valid output. This observation further indicates that models without fine-tuning exhibit limited instruction-following capabilities for the FMD task.

## 5 Conclusions and Future Work

The paper presents a technical solution to the Financial Misinformation Detection Challenge, combining retrieved high-quality evidence with a financial Chain-of-Thought (CoT) prompt to enhance prediction accuracy. However, the proposed pipeline demonstrates limitations in explanation quality compared to fine-tuned baselines. This emphasizes the necessity of incorporating fine-tuning steps to improve performance in future work.

## Limitation

Due to limited computing resources, the open-source models used in this study are restricted to the 7B/8B parameter scale. Additionally, our method has not undergone a fine-tuning step, and the retrieved results are relatively sparse. In next step, we will involve fine-tune step to further analysis the effectiveness of Financial CoT and we aim to extract key information more effectively from broader network search results to better support prediction.

## References


Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. *arXiv preprint arXiv:2106.05707*.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2023. Complex claim verification with evidence retrieved in the wild. *arXiv preprint arXiv:2305.11859*.

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.

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.

Betty Liu and Austin Moss. 2022. The role of accounting information in an era of fake news. *Available at SSRN 4399543*.

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*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.13379*.

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

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. 2024. Evidence-driven retrieval augmented response generation for online misinformation. *arXiv preprint arXiv:2403.14952*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xichen Zhang and Ali A Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2):102025.


## A Appendix

### A.1 Zero-Shot Prompt

Zero-Shot Prompt.

**System Message:** You are a Fact Checker and you need to focus on the financial sector. Given a claim, assess the factual accuracy of the claim based on the evidence and generate the explaination. Make a prediction and provide an explanation for your prediction.

**User Message:** I will give you one claim and relevant evidence. Your task is to verify the factual authenticity of the claim based on the evidence provided. Make a final prediction from: 'True', 'False' or 'Not Enough Info' and provide a detailed explanation. Please provide the final output in JSON format containing the following two keys: prediction and explanation.