Deloitte (Drocks) at the Financial Misinformation Detection Challenge Task: Enhancing Misinformation Detection through Instruction-Tuned Models

Harika Abburi[1], Alex Chandler[2], Edward Bowen[2], Sanmitra Bhattacharya[2], Nirmala Pudota[1]

[1]Deloitte & Touche Assurance and Enterprise Risk Services India Private Limited India
[2]Deloitte & Touche LLP, USA
{abharika, achandler, edbowen, sanmbhattacharya, npudota}@deloitte.com

Abstract

Large Language Models (LLMs) are capable of producing highly fluent and convincing text; however, they can sometimes include factual errors and misleading information. Consequently, LLMs have emerged as tools for the rapid and cost-effective generation of financial misinformation, enabling bad actors to harm individual investors and attempt to manipulate markets. In this study, we instruction-tune Generative Pretrained Transformers (`GPT-4o-mini`) to detect financial misinformation and produce concise explanations for why a given claim or statement is classified as misinformation, leveraging the contextual information provided. Our model achieved fourth place in Financial Misinformation Detection (FMD) shared task with a micro $F1$ score of 0.788 and a ROUGE-1 score of 0.743 on the private test set of FACT-checking within the FINancial domain (FIN-FACT) dataset provided by the shared task organizers.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in understanding and generating human language, particularly through the application of in-context learning (ICL) across a range of tasks and model sizes (Dong et al., 2024; Agarwal et al., 2024; Bertsch et al., 2024). With the widespread availability of LLMs, users can tackle diverse tasks simply by providing instructions, with or without examples, allowing the LLM to generate the required output.

However, while LLMs enable users to solve tasks without needing technical expertise, they also present significant risks. Malicious actors can misuse these models to generate misleading or harmful content (Andriushchenko et al., 2024b), with misinformation produced by LLMs often being more challenging to detect than that authored by humans (Chen and Shu, 2024). As research advances in aligning language models to user intentions and preventing misuse, efforts to bypass these safeguards, known as jail-breaking, have also intensified (Chao et al., 2024). Despite the implementation of guardrails, certain strategies can circumvent the safety measures of state-of-the-art (SOTA) LLMs (Andriushchenko et al., 2024a). Additionally, numerous fine-tuned LLMs may lack acceptable safeguards, making them vulnerable to harmful instructions (Chan et al., 2023; Qi et al., 2023; Henderson et al., 2024).

One of the concerning forms of harmful content is misinformation (or false or misleading information), with (Thibault et al., 2024) identifying at least 75 distinct types covering health, politics, celebrities, rumors, and deepfakes. In the financial domain, misinformation is particularly harmful, as it has the potential to disrupt markets and negatively impact investors by spreading false information about financial products or companies (Rangapur et al., 2023b). Given the rapid, cost-effective production of misinformation, coupled with the time-intensive process of manual verification, there is an urgent need to automate the detection and flagging of misinformation. Such automation should not only correctly identify false information but also provide clear explanations of the factors that make the content misleading.

Misinformation detection approaches include rule-based methods with keyword analysis and heuristic rules (Papageorgiou et al., 2024), traditional deep learning methods and pre-trained models (Kamal et al., 2023; Chung et al., 2023; Rangapur et al., 2024), and LLMs or Vision Language Models (VLMs) (Alghamdi et al., 2024). However, as observed by (Liu et al., 2024), the pre-trained models exhibit poor performance in detecting financial misinformation, likely due to their smaller parameter sizes limiting their ability to comprehend long, complex texts and subtle forms of misinformation. The two most actively researched frame-

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 313–320
January 19–20, 2025. ©2025 Association for Computational Linguistics

works for misinformation detection are LLM-based frameworks (Whitehouse et al., 2022; Wan et al., 2024; Hu et al., 2024; Wu et al., 2024) and multimodal frameworks, often including VLMs (Abdelnabi et al., 2022; Wang et al., 2024; Qi et al., 2024).

The exploration of LLM-based methods for detecting financial misinformation has become a prominent area of research. To boost this further, Financial Misinformation Detection (FMD) organizers[1] introduced a task aimed at detecting financial misinformation with concise explanations. In this work, we instruction-tuned (IT) `GPT-4o-mini` (referred as *GPT-4o-mini-IT* in rest of the paper) to classify news headlines in the FACT-checking within the FINancial domain (FIN-FACT) dataset (Rangapur et al., 2023a), providing labels (True, False, Not Enough Information) and explanations justifying the classification of claims. Our experiments show that our instruction-tuned model outperforms several baselines using well established evaluation metrics.

2 FIN-FACT Dataset

FIN-FACT dataset (Rangapur et al., 2023a) is a multimodal benchmark dataset to evaluate financial fact-checking of claims. It contains claims from diverse financial sectors such as Income, Finance, Economy, Budget, Taxes, and Debt, and with labels assigned as 'True', 'False', and 'NEI' (Not Enough Information) according to the provided justification. The dataset is carefully designed to reflect the complexity of financial narratives by including contextual information, supporting evidence links, and visual elements such as image links and captions for each claim. A notable feature of this dataset is the availability of explanations justifying the classification of each claim. This feature significantly enhances its value for training language models to not only detect misinformation but also generate well-reasoned explanations for their evaluations.

The dataset contains the following columns:

- **claim**: core assertion
- **posted date**: temporal information
- **sci-digest**: claim summaries
- **justification or context**: offers insights to further contextualize claim
- **image link**: visual information
- **issues**: claim complexities
- **label**: 'True' or 'False' or 'NEI'
- **evidence**: ground truth explanations

Label	Number of training samples	Number of validation samples
True	642	75
False	809	83
NEI	306	38
Total	1757	196

Table 1: FIN-FACT dataset statistics

To enable analysis of the claims, we introduced an **updated_claim** column by concatenating the 'claim' and 'sci-digest' fields. The claim column often contained only a few words, while the 'sci-digest' column provided detailed information. This combination ensures the model receives more specific details for fact-checking. If the 'sci-digest' contained NaN values, we bypassed the concatenation and used the claim data as it was.

Upon manual inspection, we identified that many image URLs were broken, numerous claims missing associated images, and the available images often contained irrelevant information. As a result, we decided to exclude the image link column entirely. In our study, in addition to the 'updated_claim' column we created, we considered 'context', 'label', and 'evidence' columns from the FIN-FACT dataset.

Table 1 shows the distribution of samples in the training and validation sets. A subset of training samples are used to instruction-tune the GPT-4o-mini model. The shared task organizers evaluated the performance of the submissions on a test set of 1304 samples. This test set is further split into private and public subsets. The distribution of samples for each subset is not disclosed to the participants during the result submission phase. Additional details about the task and dataset are available at [1].

3 *GPT-4o-mini-IT* as a Misinformation Detector

While LLMs have been widely applied to various Natural Language Generation (NLG) tasks, their use in detecting misinformation with robust reasoning remains underexplored. We chose `GPT-4o-mini` for its SOTA zero-shot classification

[1] https://coling2025fmd.thefin.ai/home

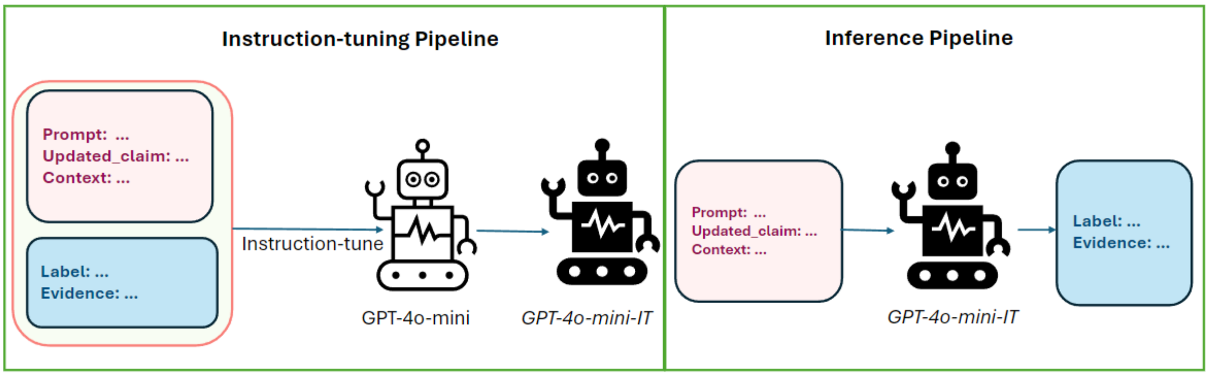

Figure 1: Our end-to-end instruction-tuning and inference pipeline

abilities and lower fine-tuning costs compared to `GPT-4o` (OpenAI, 2024b; Rahaman et al., 2024). Figure 1 presents our end-to-end instruction-tuning and inference pipeline.

Our instruction-tuning pipeline enhances `GPT-4o-mini`'s ability to detect misinformation in the financial domain and provide clear evidence. Taking advantage of its generalization capabilities, the model efficiently applies learned patterns to new claims with minimal instruction-training on only 918 samples (consisting of 306 NEI samples and an equal number for the True and False labels to create a balanced set). The model is instruction-tuned to perform a dual task: determining the truthfulness of the claim and generating a succinct explanation for the classification.

Let uc_i and co_i represent the inputs for the updated_claim and context respectively, while the ground truth label l_i and evidence e_i serve as the outputs. We perform instruction-tuning on `GPT-4o-mini` by concatenating the prompt (p), inputs (uc_i, co_i), and outputs (l_i, e_i) into a single input sequence as shown in the following message, obtaining the *GPT-4o-mini-IT* model.

message_i: [
{"role": "system", "content": "p"},
{"role": "user",
"content": "*claim: {uc_i}, context: {co_i}*"},
{"role": "assistant",
"content": "*label: {l_i}, evidence: {e_i}*"}
]

During inference, we provide the prompt, updated_claim, and context as a single input sequence to *GPT-4o-mini-IT* to generate the output (o_i), where $o_i = (l_i, e_i)$. The output o_i is then post-processed to extract the label and evidence, where $l_i \in$ {True, False, NEI} and e_i represents the explanation justifying the classification.

3.1 Choice of Prompt and Experimental Settings

During the development of the system prompt, we performed detailed prompt engineering to determine the suitable prompt. The final prompt (p) details are available in Appendix Section A.

To decrease variance in output, we set the *temperature* parameter to 0. We operated with a *batch size* of 3 and conduct 3 *training epochs* to allow for stability and reliability in model performance.

4 Experiments

We reported model's performance using well established metrics, namely the `micro F1 score` (`F_micro`) for ternary misinformation classification, and the `ROUGE-(1,2, and L)` scores (Lin, 2004) which are used to assess the quality of reasoning and evidence generated by the model. The average of `F_micro` and `ROUGE-1` is taken as the final ranking metric (Overall) in the challenge. We therefore used the same metric to provide a fair comparison.

4.1 Baselines

To establish a strong baseline, we explored both open-source and proprietary LLMs. We applied zero-shot prompting using the same prompt (as mentioned in Appendix Section A) on the following LLMs: `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` (OpenAI, 2023) and `GPT-4o-mini` (OpenAI, 2024a).

4.2 Results

Table 2 shows the performance of the instruction-tuned *GPT-4o-mini-IT* model compared to other

Model	Overall	F_micro	ROUGE-1	ROUGE-2	ROUGE-L
Vicuna-7b	0.309	0.469	0.148	0.067	0.108
Mistral-7b-Instruct	0.491	0.658	0.324	0.153	0.208
LLaMA2-chat-7b	0.494	0.653	0.336	0.157	0.204
LLaMA3-8b-Instruct	0.492	0.648	0.335	0.159	0.211
ChatGPT (gpt-3.5-turbo)	0.496	0.668	0.324	0.159	0.212
GPT-4o-mini	0.492	0.665	0.319	0.108	0.173
Our model *(GPT-4o-mini-IT)*	**0.751**	**0.776**	**0.726**	**0.684**	**0.700**

Table 2: Results on validation set with various LLMs in a zero-shot setting and our model

Model	Overall	F_micro	ROUGE-1	ROUGE-2	ROUGE-L
FMDLlama (Liu et al., 2024)	0.609	0.761	0.456	0.354	0.382
ChatGPT (gpt-3.5-turbo)	0.515	0.763	0.267	0.102	0.166
Our model *(GPT-4o-mini-IT)*	**0.788**	**0.828**	**0.748**	**0.708**	**0.723**

Table 3: Results on public test set with baselines and our model

Model	Overall	F_micro	ROUGE-1	ROUGE-2	ROUGE-L
FMDLlama (Liu et al., 2024)	0.584	0.718	0.450	0.346	0.374
ChatGPT (gpt-3.5-turbo)	0.481	0.701	0.261	0.099	0.163
Our model *(GPT-4o-mini-IT)*	**0.765**	**0.788**	**0.743**	**0.698**	**0.714**

Table 4: Results on private test set with baselines and our model

LLMs operating in a zero-shot setting on the validation dataset. Additionally, we also performed instruction-tuning on open-source LLMs; however the results were suboptimal, and therefore, we omitted them from this report.

GPT-4o-mini-IT model demonstrates notable improvements across the evaluated metrics. This instruction-tuned model achieves the highest overall score of 0.751, outperforming other models like `GPT-4o-mini` and `LLaMA` variants. The improvement in the F_micro score 0.776 highlights the model's enhanced accuracy in classifying misinformation, showcasing the benefits of instruction-tuning on specialized tasks and its robustness in addressing complex financial misinformation detection tasks.

Moreover, the improved ROUGE scores (ROUGE-1: 0.726, ROUGE-2: 0.684, ROUGE-L: 0.700) indicate that the model generates high-quality explanations, which are essential for understanding and verifying claims. While other LLMs in a zero-shot setting offer valuable baseline performance, the effectiveness of *GPT-4o-mini-IT* highlights the benefits of fine-tuning models on specific datasets.

Table 3 and 4 show the final results on public and private test sets respectively. The results on both test sets consistently highlight the significant performance of the *GPT-4o-mini-IT* model compared to other baseline models, including `FMDLlama` (an instruction-tuned version of `LLaMA3-8b-Instruct`) and `GPT-3.5-turbo` which is tested in a zero-shot setting. Our model achieved overall score of 0.788 on private test set securing fourth place in FMD competition. The results on private test set are provided on leaderboard[2].

5 Conclusion

In this study, we demonstrated that instruction-tuning `GPT-4o-mini` on a smaller dataset, significantly enhances its capability to detect misinformation with reasoning in the financial domain. Our approach outperforms previous solutions and other open-source LLMs in zero-shot settings, achieving a top-4 ranking on the FMD shared task leaderboard. As part of future work, we plan to integrate the VLMs to address the loss of visual information in our text-only framework. Additionally, we aim to investigate agent-based methods for financial misinformation detection and examine the model's multilingual capabilities to enhance the generalizability and robustness of our approach.

[2]https://coling2025fmd.thefin.ai/leaderboard. our team name is shown as *Drocks* in the leaderboard

References

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14920–14929.

Rishabh Agarwal, Avi Singh, Lei M. Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. Many-shot in-context learning. *Preprint*, arXiv:2404.11018.

Jawaher Alghamdi, Suhuai Luo, and Yuqing Lin. 2024. A comprehensive survey on machine learning approaches for fake news detection. *Multimedia Tools and Applications*, 83(17):51009–51067.

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2024a. Jailbreaking leading safety-aligned llms with simple adaptive attacks. *Preprint*, arXiv:2404.02151.

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Eric Winsor, Jerome Wynne, Yarin Gal, and Xander Davies. 2024b. Agentharm: A benchmark for measuring harmfulness of llm agents. *Preprint*, arXiv:2410.09024.

Amanda Bertsch, Maor Ivgi, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. 2024. In-context learning with long-context models: An in-depth exploration. *Preprint*, arXiv:2405.00200.

Alan Chan, Ben Bucknall, Herbie Bradley, and David Krueger. 2023. Hazards from increasingly accessible fine-tuning of downloadable foundation models. *Preprint*, arXiv:2312.14751.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2024. Jailbreaking black box large language models in twenty queries. *Preprint*, arXiv:2310.08419.

Canyu Chen and Kai Shu. 2024. Can llm-generated misinformation be detected? *Preprint*, arXiv:2309.13788.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

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

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. *Preprint*, arXiv:2301.00234.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara

Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Peter Henderson, Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, and Prateek Mitta. 2024. Safety risks from customizing foundation models via fine-tuning. Policy Brief, HAI Policy & Society.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22105–22113.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Ashraf Kamal, Padmapriya Mohankumar, and Vishal Kumar Singh. 2023. Financial misinformation detection via roberta and multi-channel networks. In *Pattern Recognition and Machine Intelligence: 10th International Conference, PReMI 2023, Kolkata, India, December 12–15, 2023, Proceedings*, page 646–653, Berlin, Heidelberg. Springer-Verlag.

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. *Preprint*, arXiv:2409.16452.

OpenAI. 2023. Gpt-4 technical report.

OpenAI. 2024a. Gpt-4o mini: Advancing cost-efficient intelligence. Accessed: 2024-07-18.

OpenAI. 2024b. Openai api pricing. Accessed: 2023-10-01.

Eleftheria Papageorgiou, Christos Chronis, Iraklis Varlamis, and Yassine Himeur. 2024. A survey on the use of large language models (llms) in fake news. *Future Internet*, 16:298.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. *Preprint*, arXiv:2403.03170.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *Preprint*, arXiv:2310.03693.

Ananya Rahaman, Anny Zheng, Mostafa Milani, Fei Chiang, and Rachel Pottinger. 2024. Evaluating sql understanding in large language models. *arXiv e-prints*, pages arXiv–2410.

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

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.

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

Camille Thibault, Gabrielle Peloquin-Skulski, Jacob-Junqi Tian, Florence Laflamme, Yuxiang Guan, Reihaneh Rabbany, Jean-François Godbout, and Kellin Pelrine. 2024. A guide to misinformation detection datasets. *Preprint*, arXiv:2411.05060.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. Dell: Generating reactions and explanations for llm-based misinformation detection. *Preprint*, arXiv:2402.10426.

Jiazhen Wang, Bin Liu, Changtao Miao, Zhiwei Zhao, Wanyi Zhuang, Qi Chu, and Nenghai Yu. 2024. Exploiting modality-specific features for multi-modal manipulation detection and grounding. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4935–4939.

Chenxi Whitehouse, Tillman Weyde, Pranava Madhyastha, and Nikos Komninos. 2022. Evaluation of fake news detection with knowledge-enhanced language models. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1425–1429.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3367–3378.

A Appendix

Our Financial Misinformation Detection Prompt

****Role:****

Senior Financial Misinformation Detection Specialist.

****Objective:****

Evaluate the truthfulness of financial claims with precision and substantiate your conclusions with compelling evidence.

Instructions:
1. **Input Details:**
You will be provided with two integral components for each analysis task - a Claim and its corresponding Context

2. **Assessment Process:**
- Begin with a close and thorough reading of both the Claim and the Context to grasp the full scope of information.
- Analyze the relationship between the Claim and the Context by considering the following categories:
- **True**: Assign this label under these conditions:
- The Context contains clear, unambiguous evidence that directly confirms the Claim.
- Each element within the Context consistently aligns to support the entire Claim without any need for conjecture.
- **False**: Utilize this label when:
- The Context includes specific information that clearly refutes any aspect of the Claim.
- Contradictions are apparent and do not require external analysis or interpretation.
- **Not Enough Information (NEI)**: Use NEI if:
- The Context lacks the necessary detail or completeness to unequivocally determine the Claim's accuracy or inaccuracy.
- Ambiguities, data gaps, or indirect references prevent a conclusive decision.
- Any necessity for assumptions or external context to affirm the Claim extends beyond the provided details.

3. **Evidence Compilation:**
Upon determining the label, distill and document explicit and pertinent evidence from the Context that underpins your conclusion. Prioritize evidence that decisively influences your decision to ensure clarity and coherence.

Output Requirements:
- **Predicted Label:** Clearly state your conclusion with one of the following labels: "True," "False," or "NEI."
- **Supporting Evidence:** Concisely summarize and list all significant evidence from the Context that corroborates your Predicted Label, ensuring each piece directly relates to the Claims being evaluated.

Additional Considerations:
- Employ a systematic, step-by-step reasoning approach to ensure no detail is missed during evaluation.
- Exercise critical thinking and scrupulously verify facts before finalizing your judgment.
- Aim for impartiality, accuracy, and clarity in both your analysis and the presentation of your supporting evidence.