Team FMD LLM at the Financial Misinformation Detection Challenge Task: Exploring Task Structuring and Metadata Impact on Performance

Ken Kawamura
Independent Scholar
ken_kawamura@alumni.brown.edu

Abstract

The detection of financial misinformation (FMD) is a growing challenge. In this paper, we investigate how task structuring and metadata integration impact the performance of large language models (LLMs) on FMD tasks. We compare two approaches: predicting the label before generating an explanation, and generating the explanation first. Our results reveal that prediction-first models achieve higher F1 scores. We also assess the effect of auxiliary metadata, which surprisingly degraded performance despite its correlation with the labels. Our findings highlight the importance of task order and the need to carefully consider whether to use metadata in limited data settings.

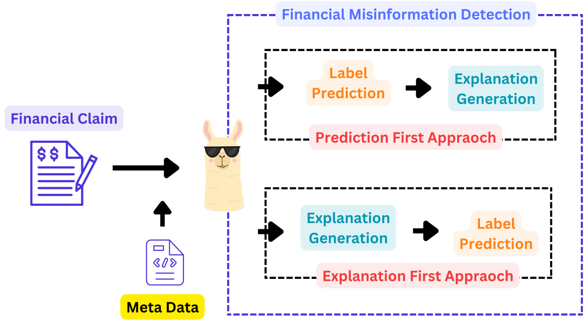

Figure 1: Overview

1 Introduction

Recently, Large Language Models (LLMs) (Sanh et al., 2021; Brown et al., 2020; Achiam et al., 2023; Scao et al., 2022; Touvron et al., 2023) has been transforming finance sectors with their adaptation (Shah et al., 2022; Wu et al., 2023; Xie et al., 2023; Kawamura et al., 2024). At the same time, there is a growing need to automate the detection of misinformation in finance, where misinformation can lead to market manipulation and instability (Rangapur et al., 2023b; Mohankumar et al., 2023; Chung et al., 2022; Liu et al., 2024).

In this paper, we present our approach to the Financial Misinformation Detection (FMD) shared task at COLING 2025, where we developed models capable of both classifying financial claims and generating explanations for the predictions. Our experiments revealed two key insights: (1) classifying claim labels prior to generating explanations significantly improved classification performance in F1 score, challenging the common practice of generating reasoning as a precursor to prediction, such as in Chain of Thought prompting; and (2) incorporating auxiliary metadata, such as summary fields, unexpectedly degraded model performance, despite the strong correlation of this metadata with the labels. This finding challenges conventional assumptions about feature engineering, in tasks requiring nuanced reasoning with limited data.

2 Related Studies

The growing interest in fact-checking spans various domains, from addressing misinformation related to COVID-19 (Saakyan et al., 2021), to verifying health-related claims (Sarrouti et al., 2021), to checking scientific assertions (Wadden et al., 2020), and even to creating large-scale, multi-domain datasets such as FEVER (Thorne et al., 2018). In the financial domain, the detection of misinformation has emerged as an important focus. For example, Rangapur et al., 2023a introduced the Fin-Fact dataset, specifically designed to address the gap in domain-specific fact-checking resources for financial misinformation.

Earlier research in financial misinformation detection primarily utilized traditional NLP techniques, including RoBERTa (Liu et al., 2019), LSTM-based models, and custom neural architectures (Kamal et al., 2023; Chung et al., 2022; Mohankumar et al., 2023). With increasing evaluations of LLMs in fields like the legal domain (Stern et al., 2024), there is a growing need for similar assessments in financial misinformation detection.

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 288–296
January 19–20, 2025.

Recent advancements, particularly the work by Liu et al., 2024, have leveraged domain-specific fine-tuning for LLMs. Their fine-tuned version of `llama3.1-8b`[1] outperformed leading zero-shot models, such as `Mistral-7b-Instruct` (Jiang et al., 2023) and `Gemma-instruct-7b` (Mesnard et al., 2024), highlighting the benefits of fine-tuning LLMs over general-purpose models in financial misinformation detection.

3 Task and Dataset

3.1 Task Description

The Financial Misinformation Detection (FMD) task is a multitask learning challenge where models classify financial claims into three categories—True, False, or Not Enough Information (NEI)—and generate explanations for their classifications. This dual objective emphasizes accurate classification and the interpretability of the model's predictions, ensuring they are substantiated by relevant financial evidence. Task organizers encourage fine-tuning large language models (LLMs) and prompt engineering.

3.2 Dataset

Participants were provided with 1,953 labeled training examples and 1,304 test examples from the Fin-Fact dataset (Rangapur et al., 2023a)[2], which includes fields such as *claim*, *label* (True, False, NEI), *explanation*, and *justification*. The *label* indicates the veracity of the claim, while the *explanation* provides a free-form textual rationale supporting the assigned label. *Justifications* offer additional arguments in favor of the claims. To further enrich this context, additional metadata—such as the *posting date*, *image*, and *sci_digest* summaries (i.e., brief claim overviews)—were included. However, some metadata fields, like *sci_digest*, were not always available and could be empty. A baseline prompt was also provided by the organizers to guide initial model development[3] (Appendix A). Table 1 presents sample entries from the dataset.

3.3 Data Exploration

To gain deeper insights into the dataset, we conducted an exploratory analysis of the provided metadata fields. One notable finding emerged: cases where the *sci_digest* field was absent were highly correlated with the True label (339 out of 364 instances). Building on this observation, we developed a heuristic: if the *sci_digest* field is empty, the label is predicted as True; otherwise, the label is predicted as False. Applied to the training data, the heuristic achieved an F1 of 62.1%, surpassing the random baseline's 34.2%, showcasing the potential of metadata-driven approaches (Appendix C).

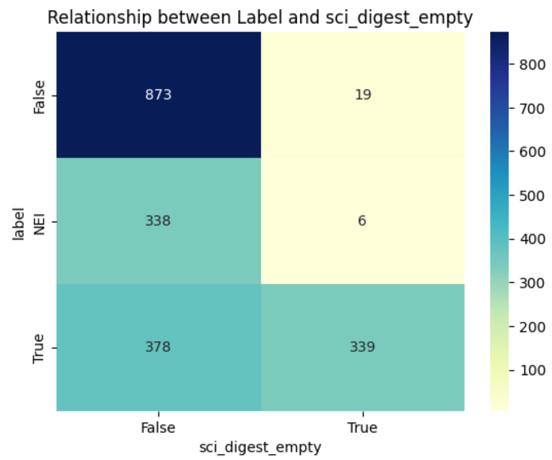

Figure 2: Relationship between label and whether *sci_digest* is empty

We examined other metadata, such as image metadata availability, but *sci_digest* showed the strongest label correlation. Its binary nature suited simple feature engineering, while richer metadata like temporal or visual data is left for future work.

However, the availability of the *sci_digest* field should not determine a claim's veracity. Whether the field is present or empty—merely reflecting data collection artifacts—does not provide meaningful insight into the claim's truth. For example, reasoning that a claim is True because the *sci_digest* field is empty is a superficial pattern, not a valid explanation. The heuristic's success stems from this pattern, not from any real contribution to misinformation detection.

4 Approach

Our approach optimized financial misinformation detection by developing prompts tailored to two key factors: (1) subtask order, comparing whether classifying a financial claim (True/False/NEI) before generating an explanation yields better performance than the reverse, and (2) the potential benefits of leveraging auxiliary metadata, particularly the availability of *sci_digest* field, which showed strong label correlations.

[1]https://www.llama.com/

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

[3]https://github.com/lzw108/COLING25-FMD/blob/main/practice_data_preprocess.ipynb

Table 1: Examples of claims, labels, and corresponding explanations from the Fin-Fact dataset.

Label	Claim	Explanation
True	Tax rates were significantly higher in the '40s, the '50s, and the '60s.	Today, tax rates range from 10 percent for lower incomes to 35 percent for the highest incomes. (See a chart of tax rates over timefrom the Tax Foundation here.)
False	Texas this fiscal year will have more money in reserve than the other 49 states combined.	In the Feb. 25, 2015 interview,which we caught online, Patrick said: We are in the best financial shape of any state in the country. Well have about $11 billion or so in our rainy day fund by the end of our fiscal year. ...
NEI	Beto O'Rourke's 'Reality Check' can be paraphrased as "A thorough evaluation of the facts by Beto O'Rourke."	One such meme, entitled "'Beto' Reality Check," was shared widely on Facebook in August 2018:A spokesperson for O'Rourke's campaign described the meme as "factually incorrect in countless ways" and largely referred us to several existing news reports about the allegations. The following is our breakdown of the five sections contained in the meme.O'Rourke adopted the name "Beto" to appeal to Latino voters:...

To evaluate these aspects, we fine-tuned `Llama-3.2-1B-Instruct`[4]. We hypothesized that in a complex task with limited training data, such as the FMD, both subtask order and metadata inclusion could significantly impact model performance.

4.1 Baselines

We adopted the baseline study by Liu et al., 2024, which evaluated multiple LLMs using the challenge organizers' baseline prompt, including `ChatGPT` (`gpt-3.5-turbo`) and FMDLlama (Liu et al., 2024), a model fine-tuned for the FMD task.

4.2 Generation Order

Chain of Thought prompting, where a model generates an intermediate reasoning process before arriving at a final answer, is a common technique for improving model reasoning (Wei et al., 2022). We hypothesized that generating the explanation first, rather than producing it post hoc, could similarly enhance the model's performance. By generating the explanation upfront, the model can fully evaluate the claim before classifying it, potentially improving prediction accuracy as the reasoning unfolds.

Conversely, predicting the label first may simplify the task for the model. Since the labels (True, False, NEI) are fixed, the output always begins with one of these three options, making the task more structured. In contrast, generating the explanation first adds complexity, as the model must not only generate coherent reasoning but also determine when to stop reasoning, and transition to classification. The label-first approach might better optimize the classification task by making the problem straightforward for LLMs to learn, especially when training data is limited as in the FMD task.

[4] `https://huggingface.co/unsloth/Llama-3.2-1B-Instruct-bnb-4bit`

```
Please determine whether the claim is True, False, or Not Enough Information (NEI) based on contextual information, and provide an appropriate explanation. The answer needs to use the following format:
Prediction: [True, or False, or NEI]
Explanation: [Explain why the above prediction was made]
### Claim:
{claim}

### Contextual Information
{justification}

### Prediction:
{True, False, or NEI}

### Explanation:
{explanation}
```

Figure 3: Prompt for Prediction First Without Metadata

4.3 Auxiliary Metadata

Incorporating auxiliary metadata that correlates with target labels can enhance prediction accuracy by allowing the model to exploit known patterns. For example, our analysis of *sci_digest* field revealed a strong correlation between its absence and the True label. Including this metadata in the prompt could help the model exploit these correlations, improving its predictions without requiring deep semantic understanding.

However, the presence or absence of the *sci_digest* field does not provide semantic insight into claim veracity. Its utility stems from superficial data patterns. Large language models, designed to reason through typical natural language inference patterns, may struggle to leverage metadata-driven patterns that lack explicit linguistic meaning. This limitation could hinder the model's ability to generate accurate predictions when relying too heavily on metadata like whether *sci_digest* field is empty.

4.4 Prompt Design

To assess the impact of generation order and metadata inclusion, we designed prompts with varying structures. In one version, the model predicted the claim's label (True/False/NEI) before generating

Model	Overall Score	*Classification* Micro-F1	*Explanation* ROUGE-1	ROUGE-2	ROUGE-L
Baselines					
`ChatGPT (gpt-3.5-turbo)`	0.5152	0.7634	0.267	0.102	0.1662
`FMDLlama`	0.6089	**0.7616**	0.4563	0.3536	0.3817
Ours					
`Prediction First (No Metadata)`	**0.6285**	0.7357	**0.5213**	0.4487	0.4683
`Explanation First (No Metadata)`	0.5631	0.6063	0.5200	**0.4501**	**0.4667**
`Prediction First (With Metadata)`	0.5914	0.6969	0.4860	0.4150	0.4340
`Explanation First (With Metadata)`	0.5086	0.4972	0.5199	0.4495	0.4669

Table 2: Performance of different models and prompt configurations on the public test set of the FMD task. Results for the private test set, where only one model was allowed for evaluation, are detailed in Appendix D.

Please determine whether the claim is True, False, or Not Enough Information (NEI) based on contextual information, and provide an appropriate explanation. The answer needs to use the following format:
Explanation: [Explain why the above prediction was made]
Prediction: [True, or False, or NEI]
Claim:
{claim}

Contextual Information
{justification}

sci_digest is empty:
{True or False}

Explanation:
{explanation}

Prediction:
{True, False, or NEI}

Figure 4: Prompt for Explanation First With Metadata

an explanation, while in another, the explanation was generated first. Additionally, we evaluated the influence of metadata by creating two types of prompts: one that incorporated the *sci_digest* field and another that excluded it. Figure 3 illustrates the Prediction First approach without Metadata, while Figure 4 showcases the Explanation First approach with Metadata, including the handling of the *sci_digest* field.

4.5 Model Fine-tuning

We finetuned `Llama-3.2-1B-Instruct` in 4 bit using Unsloth[5]. We trained a model per prompt template for three epochs, and they all had the best validation loss at the end of three epochs. The detailed hyperparameters can be found in the Appendix B.

5 Results

Table 2 presents the performance on the public test set of different models and prompt configurations across key metrics: micro-F1-score, ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004). The overall score for this task was computed as average of F1 and ROUGE-1. Out best model (Overall Score: 0.6285) outperformed both ChatGPT (Overall Score: 0.5152) and FMDLamma (Overall Score: 0.6089). More importantly, the results highlight the impact of task order (classification prediction before explanation vs. explanation before classification prediction) and the inclusion of metadata on model performance.

Our findings indicate that models predicting the label before generating an explanation achieve higher F1 scores. Prediction First without Metadata (Micro-F1: 0.7357) performed better than Explanation First without Metadata (Micro-F1: 0.6063) by 0.1294. Additionally, Prediction First with Metadata (Micro-F1: 0.6969) performed better than Explanation First with Metadata (Micro-F1: 0.4972) by 0.1997. This supports the hypothesis that beginning with the more constrained task of classification leads to better overall performance in financial misinformation detection.

Including whether *sci_digest* is empty (metadata) consistently lowered F1 scores, suggesting that while metadata correlates with labels, it may hinder model performance. Specifically, the inclusion of metadata reduced the F1 score by 0.0388 in the Prediction First approach and by 0.1091 in the Explanation First approach. This implies that metadata may need to offer more than surface-level correlations to be effective in enhancing the model's reasoning process

6 Conclusion

Our results demonstrate that predicting the label before generating an explanation improves classification performance in financial misinformation detection, as evidenced by F1 score. This contrasts with conventional approaches that prioritize reasoning-first strategies. Additionally, the inclusion of auxiliary metadata, such as the *sci_digest* field, despite its high correlation with the labels, hindered model performance. This finding challenges conventional assumptions regarding the benefits of metadata for prediction tasks, especially in cases where the metadata lacks semantic richness.

[5] `https://unsloth.ai/`

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haim ing Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Is abella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Ma teusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack W. Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin D. Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qim ing Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Wingyan Chung, Yinqiang Zhang, and Jia Pan. 2022. A theory-based deep-learning approach to detecting disinformation in financial social media. *Information Systems Frontiers*, 25:473 – 492.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/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*.

Ken Kawamura, Zeqian Li, Chit-Kwan Lin, and Bradley McDanel. 2024. Revelata at the FinLLM challenge task: Improving financial text summarization by restricted prompt engineering and fine-tuning. In *Proceedings of the Eighth Financial Technology and*

Natural Language Processing and the 1st Agent AI for Scenario Planning, pages 146–152, Jeju, South Korea. -.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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*, abs/1907.11692.

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

Gemma Team Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, L. Sifre, Morgane Riviere, Mihir Kale, J Christopher Love, Pouya Dehghani Tafti, L'eonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Am'elie H'eliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Cl'ement Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Pier Giuseppe Sessa, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vladimir Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Brian Warkentin, Ludovic Peran, Minh Giang, Cl'ement Farabet, Oriol Vinyals, Jeffrey Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *ArXiv*, abs/2403.08295.

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.

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

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

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic. *ArXiv*, abs/2106.03794.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan D. Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng-Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization. *ArXiv*, abs/2110.08207.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Conference on Empirical Methods in Natural Language Processing*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ili'c, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurenccon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa Etxabe, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris C. Emezue, Christopher Klamm, Colin Leong, Daniel Alexander van Strien, David Ifeoluwa Adelani, Dragomir R. Radev, Eduardo Gonz'alez Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady ElSahar, Hamza Benyamina, Hieu Trung Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jorg Frohberg, Josephine Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan

Dey, Manuel Romero Muñoz, Maraim Masoud, Mar'ia Grandury, Mario vSavsko, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto L'opez, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, S. Longpre, Somaieh Nikpoor, S. Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-Shaibani, Matteo Manica, Nihal V. Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Févry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiang Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Y Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre Franccois Lavall'ee, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aur'elie N'ev'eol, Charles Lovering, Daniel H Garrette, Deepak R. Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Xiangru Tang, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, S. Osher Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdenvek Kasner, Zdeněk Kasner, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ananda Santa Rosa Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ayoade Ajibade, Bharat Kumar Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David M. Lansky, Davis David, Douwe Kiela, Duong Anh Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatim Tahirah Mirza, Frankline Ononiwu, Habib Rezanejad, H.A. Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jan Passmore, Joshua Seltzer, Julio Bonis Sanz, Karen Fort, Lívia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nourhan Fahmy, Olanrewaju Samuel, Ran An, R. P. Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas L. Wang, Sourav Roy, Sylvain Viguier, Thanh-Cong Le, Tobi Oyebade, Trieu Nguyen Hai Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Kumar Singh, Benjamin Beilharz, Bo Wang, Caio Matheus Fonseca de Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel Le'on Perin'an, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Iman I.B. Bello, Isha Dash, Ji Soo Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthi Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, María Andrea Castillo, Marianna Nezhurina, Mario Sanger, Matthias Samwald, Michael Cullan, Michael Weinberg, M Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patricia Haller, Patrick Haller, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Pratap Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yashasvi Bajaj, Y. Venkatraman, Yifan Xu, Ying Xu, Yu Xu, Zhee Xao Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

Raj 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 pretrained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2335, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ronja Stern, Ken Kawamura, Matthias Stürmer, Ilias Chalkidis, and Joel Niklaus. 2024. Breaking the manual annotation bottleneck: Creating a comprehensive legal case criticality dataset through semi-automated labeling. *arXiv preprint arXiv:2410.13460*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. 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, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu 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. *ArXiv*, abs/2307.09288.

David Wadden, Kyle Lo, Lucy Lu Wang, Shanchuan Lin, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Conference on Empirical Methods in Natural Language Processing*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

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*, abs/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*, abs/2306.05443.

A Baseline Prompt

```
Please determine whether the claim is True, False, or Not Enough Information (NEI) based on contextual
information, and provide an appropriate explanation. The answer needs to use the following format:
Explanation: [Explain why the above prediction was made]
Prediction: [True, or False, or NEI]
Claim:
{claim}

Contextual Information
{justification}

Prediction:
{True, False, or NEI}

Explanation:
{explanation}
```

Figure 5: Prompt given by an organizer

B Fine-tuning Hyperparameter

We fine-tuned our models on one V100 GPU using the following hyperparameters: a per-device batch size of 8 and a gradient accumulation of 4 steps, resulting in an effective batch size of 32. The model was trained for 3 epochs with a linear learning rate scheduler initialized at 2e-4. We employed AdamW with 8-bit optimizers to reduce memory consumption and set the weight decay to 0.01.

Warmup was applied for the first 5 steps to stabilize training. FP16 precision was used. To ensure reproducibility, we used a random seed of 3407.

C Heuristic Performance in Training Set

D Leaderboard Results

Strategy	Accuracy	Precision	Recall	F1
Predict *True* if *sci_digest* empty	0.621	0.593	0.621	0.621
Random Baseline	0.342	0.382	0.342	0.342

Table 3: Performance comparison between heuristic strategy and random baseline.

Model	Overall Score	*Classification* Micro-F1	ROUGE-1	*Explanation* ROUGE-2	ROUGE-L
Baselines					
`ChatGPT (gpt-3.5-turbo)`	0.4813	0.7012	0.2614	0.0994	0.1632
`FMDLlama`	**0.5842**	**0.7182**	0.4502	0.3464	0.3743
Ours					
`Prediction First (No Metadata)`	0.5813	0.6448	**0.5178**	**0.4428**	**0.4607**

Table 4: Performance of different models on the private test set of the FMD task. Results for the other three prompt configurations are not reported, as only one final model could be submitted for evaluation on the private split, which determined the final competition rankings.