Dunamu ML at the Financial Misinformation Detection Challenge Task: Improving Supervised Fine-Tuning with LLM-based Data Augmentation

Dongjun Lee*, Heesoo Park*

Dunamu

{tonny, belle}@dunamu.com

Abstract

In this paper, we describe Dunamu ML's submission to the Financial Misinformation Detection (FMD) 2025 shared task. To address the low-resource challenge in FMD, we augmented a general domain misinformation detection dataset for training. We first collected claims, contexts, and misinformation labels from a public dataset. Then, we generated evidence for each label based on a closed LLM with few-shot examples extracted from the FMD training dataset. Finally, we oversampled the training data specific to the financial domain and augmented it with the generated data to perform supervised fine-tuning (SFT) on the LLM. When evaluated on the blind test dataset, our model achieved an F1 score of 84.67 in misinformation classification and a ROUGE-1 score of 81.21 in evidence generation, ranking first on the leaderboard in both aspects.

1 Introduction

Misinformation detection is a very important issue in this era, where information spreads quickly through social media (Chung et al., 2023). Furthermore, the evolving landscape of the application of large language models (LLMs) which often generate false information, known as "hallucination" (Huang et al., 2024), further highlights the importance of fact verification. Especially in the financial industry, the ability to discern fake news is essential for making various decisions based on information (Rangapur et al., 2023). It is crucial not only to discern whether it is fake news or not but also to have a clear understanding of the evidence behind it to make more accurate financial decisions.

Financial Misinformation Detection (FMD)[1] Challenge aims to create a specialized LLM that excels in pinpointing financial misinformation and articulating its findings. This challenge requires participants to be provided with a claim and the context related to that claim and to train a model that can both determine whether the claim is true, false, or not enough information and generate concise explanations (Liu et al., 2024).

In this work, to overcome the low-resource setting of FMD, we address the above issues by leveraging data augmentation (DA), which enriches the diversity of the dataset without constructing new data (Feng et al., 2021). We first found a public general domain dataset built on the same external resources to overcome the data deficiency of the financial sector (Yao et al., 2023). Then, we proceeded with data augmentation using a closed LLM (e.g. GPT-4). Finally, we conducted supervised fine-tuning (SFT) with the oversampled given dataset in the financial sector and the augmented dataset in the general domain.

In the experiment using the FMD 2025 hidden test set, we achieved an F1 score of 84.67 in classifying misinformation and a ROUGE-1 score of 81.21 in generating evidence, ranking first on the leaderboard in both aspects. Moreover, we demonstrated that our data augmentation method improves the performance of SFT on FMD through ablation experiments.

2 Methodology

2.1 Data Augmentation

External Data Resource To address low-resource challenges, we found public fact-checking dataset, Mocheg (Yao et al., 2023)[2] constructed from the same web sources (Snopes[3] and PolitiFact[4]). The dataset provided by the task organizer is limited to the financial domain, whereas this dataset encompasses a general domain. This data

[*] Equal contribution.

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

[2] https://github.com/VT-NLP/Mocheg

[3] https://www.snopes.com/

[4] https://www.politifact.com/

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 297–301
January 19–20, 2025. ©2025 Association for Computational Linguistics

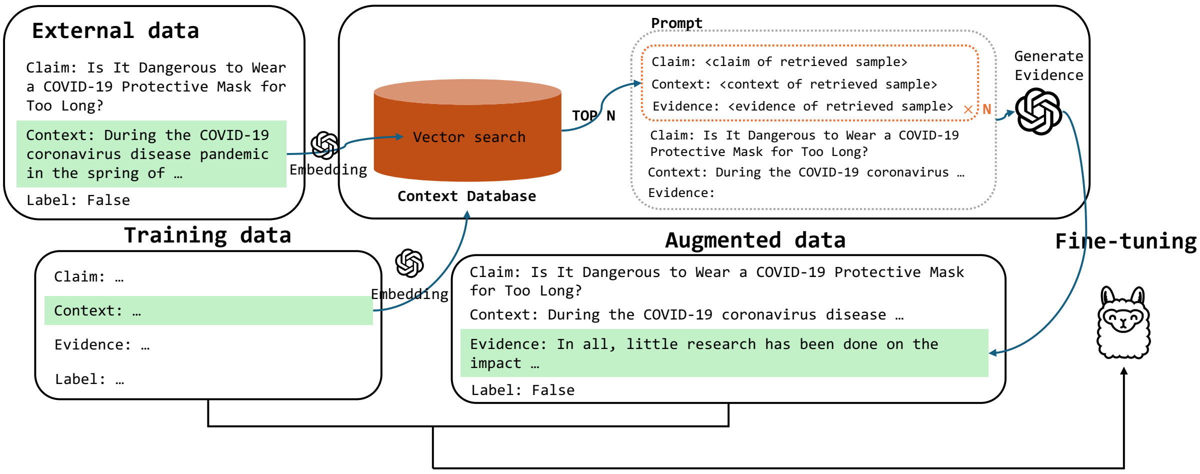
External data
Claim: Is It Dangerous to Wear a COVID-19 Protective Mask for Too Long?
Context: During the COVID-19 coronavirus disease pandemic in the spring of …
Label: False
Embedding
Vector search
Context Database
TOP N
Prompt
Claim: <claim of retrieved sample>
Context: <context of retrieved sample>
Evidence: <evidence of retrieved sample>
N
Claim: Is It Dangerous to Wear a COVID-19 Protective Mask for Too Long?
Context: During the COVID-19 coronavirus …
Evidence:
Generate Evidence
Training data
Claim: …
Context: …
Evidence: …
Label: …
Embedding
Augmented data
Claim: Is It Dangerous to Wear a COVID-19 Protective Mask for Too Long?
Context: During the COVID-19 coronavirus disease …
Evidence: In all, little research has been done on the impact …
Label: False
Fine-tuning

Figure 1: Overview of the proposed method. Our method comprises two core components: data augmentation and supervised fine-tuning.

consists of 33,880 ruling statements where each statement is mapped with a claim annotated with a truthfulness label. We automatically generated the evidence on this data using closed LLM.

Augmentation Method We applied in-context learning to generate evidence for each claim. We provide the LLM with the claim, context, and misinformation label to generate evidence, as presented in Listing 1. To generate evidence in a format similar to that in the training data, we extracted samples from the training data and provided them as few-shot examples. The criterion for selecting the few-shot examples was based on the similarity of sentence embeddings. As shown in Figure 1, for the sample for which we want to generate evidence, we selected the top-k samples from the training data with the closest context embedding similarity. Before applying augmentation, we experimented to find the appropriate closed LLM, the appropriate search key, and the number of few-shots. Detailed experimental results are presented in Section 3.3.3.

```
Generate an explanation for why a claim
    is True or False or NEI (Not Enough
    Information) based on the given
    context.
Your answer should be a part of the
    given context, meaning it should be
    extractive.

<examples>
# Claim: {example_claim}
# Context: {example_justification}
# Label: {example_label}
# Evidence: {example_evidence}

...
</examples>

Following the examples above, extract
    the evidence from the context that
    supports the label.
# Claim: {claim}
# Context: {justification}
# Label: {label}
# Evidence:
```

Listing 1: Prompt template for evidence generation.

2.2 LLM Fine-Tuning

We oversampled the given training data and combined it with the generated data for training. We performed supervised fine-tuning (SFT) (Ouyang et al., 2022) on the open-source LLM using the prompt shown in Listing 2. The LLM is fine-tuned to take a task instruction, claim, and context as input to generate a label and evidence. In other words, it is trained to generate the text that follows "# Prediction:".

```
Please determine whether the claim is
    True, False, or Not Enough
    Information (NEI) based on the given
     context, and provide appropriate
    evidence. Note that your evidence
    must be extractive from the context.

# Claim: {claim}
# Context: {justification}
# Prediction: {label}
# Evidence: {evidence}
```

Listing 2: Prompt template for supervised fine-tuning.

3 Experiments

3.1 Experimental Setup

We used only 85% of the given 1,953 training data for training, and the remaining 15% was used as the dev set. The data generated through our data

Team	F1	ROUGE-1	ROUGE-2	ROUGE-L
Dunamu ML	**84.67**	**81.21**	**78.73**	**79.69**
GGbond	79.55	78.92	75.17	76.63
1-800-SHARED-TASKS	82.83	72.53	67.63	69.11
Drocks	78.77	74.29	69.83	71.42
GMU-MU	75.75	57.89	49.56	51.45
Ask Asper	78.24	51.06	40.25	42.21
Team FMD LLM	64.48	51.78	44.28	46.07
Capybara	72.21	30.33	10.14	17.40

Table 1: The F1 and ROUGE scores for the blind test set.

Methodology	F1	ROUGE-1	ROUGE-2	ROUGE-L
only train data	83.73	79.06	75.99	77.17
only generated data	83.33	53.32	44.62	47.15
gpt-4	74.80	56.37	47.95	50.40
train data + generated data (ours)	**85.37**	**79.37**	**76.70**	**77.78**

Table 2: Ablation study for the dev set.

augmentation process amounted to 23,546. As a final training dataset, we oversampled the FMD train dataset that consists of 1,660 samples 5 times and merged them with the generated dataset as described in . For the evaluation metrics, the classification performance for True, False, and NEI was evaluated using the Micro-F1, while the generation of evidence was assessed using the ROUGE score.

3.2 Implementation Details

In the data augmentation process, we utilized `GPT-4-0613` (OpenAI et al., 2024) as the closed language model for evidence generation. For few-shot selection, we employed OpenAI's `text-embedding-3-large` for sentence embedding and used cosine similarity as the similarity metric. Additionally, we employed the `FAISS` (Douze et al., 2024) library for conducting the embedding similarity search.

For fine-tuning, we used `Llama-3.1-8B` (Dubey et al., 2024) as the pre-trained LLM, and set the maximum sequence length to 8192. For fine-tuning, we utilized eight NVIDIA A100 80GB GPUs in a single node. We used the AdaFactor optimizer (Shazeer and Stern, 2018) with a learning rate of 3e-4 and a cosine scheduler. For parameter-efficient fine-tuning, we used QLoRA (Dettmers et al., 2024) with $r = 8$ and $\alpha = 16$. We applied early stopping with 5 epochs, and the per-device batch size was set to 2. During inference, we employed beam search decoding with a beam size of 3.

3.3 Result and Analysis

3.3.1 Main Result

Table 1 presents the F1 and ROUGE scores on the blind test set. Our proposed method achieved an F1 score of 84.67 and a ROUGE-1 score of 81.21, which are the highest scores in both F1 and ROUGE metrics on the leaderboard. This result demonstrates the effectiveness of our data augmentation and fine-tuning approach in both misinformation classification and evidence generation.

3.3.2 The Effect of Data Augmentation

To further explore the effect of data augmentation, we conducted an ablation study with the following settings: 1) fine-tuning only with the given training data, 2) generation based on GPT-4, 3) fine-tuning only with the generated dataset, and 4) fine-tuning utilizing both the given training data and the generated data, as proposed. The ablation results for the development set are presented in Table 2. When we incorporated the augmented data for fine-tuning, the F1 score improved by +1.60 and the ROUGE-1 score by +0.31 compared to using only the given training data. This validates that our data augmentation contributed to the improvement in performance. When we generated labels and evidence using GPT-4, the performance signifi-

Model	Search key	# few-shot	ROUGE-1
gpt-4	claim	2	55.48
gpt-4	claim	3	55.45
gpt-4	just_head	2	**56.37**
gpt-4	just_tail	2	55.78
gpt-4o	claim	2	42.73
gpt-4o	claim	10	50.67
gpt-4o	claim	20	53.19
gpt-4o	claim	30	51.30
gpt-4o	just_head	2	43.37
gpt-4o	just_head	20	53.34

Table 3: Evidence generation results in different settings. The “just_head” refers to the first 1,000 characters of the justification and “just_tail” refers to the last 1,000 characters of the justification.

cantly decreased compared to when we applied fine-tuning, demonstrating that our fine-tuning approach is a reasonable choice. When only the generated data was used for training, the F1 score decreased by -2.04 and the ROUGE-1 score notably decreased by -26.05 compared to our proposed method, indicating that using the given training data is essential for performance.

3.3.3 The Performance on Evidence Generation

We experimented with performance variations in generating evidence based on a closed LLM from the following three perspectives: 1) the choice of LLM, 2) features utilized for selecting few-shots, and 3) the number of few-shots. Table 3 shows the result. Despite using fewer few-shot examples due to GPT-4’s token length limit (8K), it demonstrated higher performance compared to GPT-4o. In GPT-4, the maximum number of few-shot examples we could use was 3, and there was no significant difference in performance between providing 2-shots or 3-shots. In GPT-4o, when the number of few-shots increased from 10 to 20, the ROUGE-1 score improved, but when it increased to 30, the score actually decreased. Finally, when selecting few-shot examples, it was observed that choosing samples with similar justifications resulted in better evidence generation performance than choosing samples with similar claims. Due to the prompt length limit, only the first 1000 characters or the last 1000 characters of the justification were used, and using the first resulted in better performance.

4 Conclusion

This paper describes Dunamu ML’s submissions to the FMD 2025 shared task. We proposed a data augmentation method for FMD. We collected context, claims, and misinformation labels from the general domain and generated evidence using a closed LLM. Then, we oversampled the data from the financial domain and merged it with the generated data from the general domain. Finally, we performed supervised fine-tuning of the LLM using this merged dataset. When evaluated on the hidden test set, our model has achieved the top position on the leaderboard in both misinformation classification and evidence generation.

References

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

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

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

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

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, 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, Haiming Bao, Mohammad 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, Ben 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, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha 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, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz 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, Mateusz 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 Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario 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 Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll 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, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2733–2743. ACM.