

FIN-FACT: A Benchmark Dataset for Multimodal Financial Fact Checking and Explanation Generation

Aman Rangapur¹ Haoran Wang¹ Ling Jian² Kai Shu¹

¹Illinois Institute of Technology, Chicago, IL, USA

²School of Economics and Management, China University of Petroleum, Qingdao, China
{arangapur, hwang219}@hawk.iit.edu, bebetter@upc.edu.cn, kshu@iit.edu

Abstract

Misinformation can have detrimental effects on various aspects of our society. The recent incident, where a cryptocurrency exchange company failed, has led to substantial losses for investors. There is compelling evidence indicating that the widespread dissemination of financial misinformation played a crucial role in this situation. Recognizing the lack of benchmark datasets specifically designed to tackle financial misinformation, we have developed FIN-FACT, a comprehensive benchmark dataset curated for multimodal fact-checking and explanation generation. What sets FIN-FACT apart is its inclusion of expert fact-checker annotations and detailed justifications, infusing the dataset with both expertise and credibility. FIN-FACT adopts a multimodal approach, encompassing not only textual but also visual content, thereby providing complementary information sources for a more holistic analysis of factuality. Beyond this, the dataset offers insightful explanations accompanying each fact-check, empowering users to delve into the rationale behind fact-checking decisions and it can be used for automated systems to generate explanations. This transparency not only validates the credibility of the claims but also fosters trust in the overall fact-checking process. The FIN-FACT dataset, along with our experimental codes is available at <https://github.com/IIT-DM/Fin-Fact/>.

1 Introduction

In an era characterized by the rapid spread of misinformation and the proliferation of fake news, fact-checking has emerged as a critical tool for ensuring the accuracy and reliability of information (Saakyan et al., 2021; Wadden et al., 2020; Sarroui et al., 2021). The emergence of social media platforms and the wide accessibility of multimodal content have intensified the complexities linked with verifying the accuracy of assertions (Mishra et al., 2022). Notably, the financial sec-

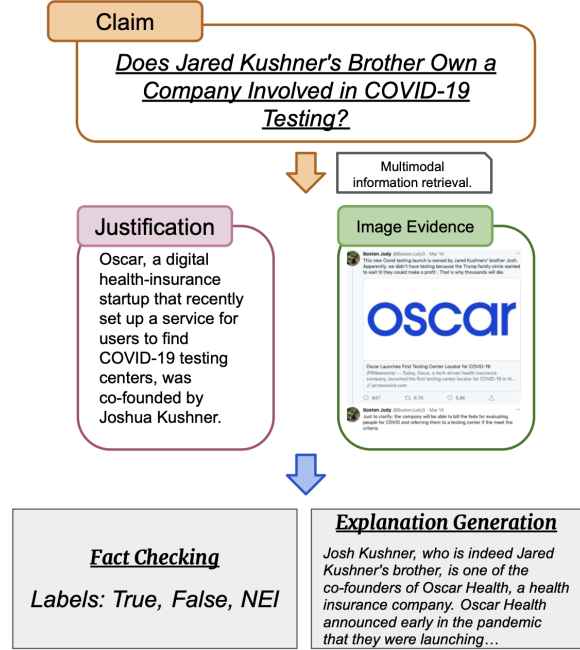


Figure 1: Illustration of comprehensive multimodal fact-checking, including True, False, and Not Enough Information (NEI), alongside the creation of explanations.

tor introduces its distinctive array of difficulties, given that precise and timely data plays a pivotal role in enabling well-informed investment choices and upholding market stability. Additionally, financial fact-checking encounters specific challenges, such as the need for customized data to address unique requirements and nuances. Furthermore, the manipulation of images to exploit visualization bias presents another significant challenge in the verification process (Mansoor and Harrison, 2018; Hüsser and Wirth, 2014).

The rise of misinformation in the financial domain has become a pressing concern, with potential impacts on public trust, investor decisions, and overall market stability (Kogan et al., 2019; Clarke et al., Forthcoming; Zhi et al., 2021; Liu and Moss, 2022). To counter the spread of misleading information, fact-checking methods have gained impor-

tance in financial reporting and analysis (Zhi et al., 2021; Mohankumar et al., 2023). However, the development of reliable models in this domain has been hindered by the lack of suitable benchmark datasets that accurately represent the intricacies of financial information and context (Rangapur et al., 2023).

Table 1: An example instance from FIN-FACT dataset.

Feature	Example
Claim	Did Biden Call for an ‘End to Shareholder Capitalism’?
Author	Dan MacGuill
Posted	08/06/2020
Sci-Digest	Fear-mongering Facebook memes misrepresented the position articulated by the presumptive Democratic presidential nominee in a speech in July 2020.
Justification	In the summer of 2020, multiple readers asked Snopes to investigate claims that former...
Evidence	In July and August, Facebook users shared posts that contained the following text...
Image	https://drive.google.com/file/1UTniyHzAB8
Issues	Debt
Label	False

In recent years, there has been notable progress in creating various datasets for fact-checking (Wadden et al., 2022; Sarrouiti et al., 2021; Saakyan et al., 2021). However, there is a noticeable gap in addressing the unique demands of fact-checking within the financial domain. Financial fact-checking faces several significant challenges. **Firstly**, it requires meticulously curated data that can encompass the intricate nuances of financial discourse. Financial documents and journalistic pieces often employ specialized language that differs from conventional structures. However, existing datasets frequently lack comprehensive coverage of financial news articles, and the absence of expert annotations diminishes the reliability of the data. **Secondly**, financial data is highly context-sensitive and constantly evolving, emphasizing the need for a dataset that can accurately capture the dynamic nature of financial markets. **Lastly**, the landscape of financial fact-checking introduces the challenge of visualization, where deliberate manipulation of visual content can shape perception and distort the accuracy of claims.

In this paper, we tackle the challenge of compiling, annotating, and refining a comprehensive corpus of financial texts that faithfully represents financial reporting, accounting methodologies, and market fluctuations. The realms of financial fact-checking and explanation generation present dis-

tinct obstacles that require specialized approaches. The necessity for tailored data capable of navigating financial terminology and intricate visual elements underscores the interdisciplinary nature inherent in this research endeavor. Figure 1 illustrates a comprehensive multimodal fact-checking and the creation of explanations while Table 1 displays an example instance from the corpus.

We introduce FIN-FACT, a new benchmark dataset created specifically for the task of multimodal financial fact-checking and explanation generation. Our contributions are as follows:

- We introduce FIN-FACT, the benchmark dataset designed to verify multimodal claims within the financial domain.
- FIN-FACT enables explanation generation through expert comments provided by fact-checking professionals.
- Our investigation show that the state-of-the-art models have difficulty performing well on FIN-FACT in the open-domain context, indicating the need to improve how these systems handle different types of data.

2 Related Work

Fact Checking. Significant efforts have been dedicated to creating fact-checking datasets for automated fact-checking systems (Wadden et al., 2020, 2022; Thorne et al., 2018; Saakyan et al., 2021). Previous studies have predominantly focused on predicting the accuracy of claims from diverse sources. While large-scale datasets from various domains have been utilized (Gupta and Srikumar, 2021), they might not be suitable for identifying misinformation related to financial matters due to domain-specific disparities. A few existing datasets on fact checking are listed in Table 2.

Although general-content misinformation datasets are readily accessible, only a limited number of datasets pertain to online financial misinformation (Clarke et al., Forthcoming; Hossain et al., 2020; Kogan et al., 2019; Zhi et al., 2021; Liu and Moss, 2022; Zhang et al., 2022; Boehm and Kroner, 2023). Current financial misinformation datasets lack clear labeling and justifications, raising concerns about result reliability. In contrast, the FIN-FACT dataset is distinct with genuine data and a multimodal structure, combining text and images to encompass

Table 2: Existing Datasets on Fact Checking.

Dataset	Topic	Size(# of instances)	Modality	Source	# of classes
LIAR	Political	12,836	Text	Politifact	6
FEVER	Multi-Domain	185,445	Text	Wikipedia	3
FakeNewsNet	Political & Celebrity	602,659	Text	Twitter	2
PHEME	Multi-Domain	330	Text	Twitter	2
MM-COVID	Medical	11,173	Text, Image	Politifact, Snopes, Poynter & Fullfact	2
Factify	Multi-Domain	50,000	Text, Image	Twitter	5
Fauxtography	Multi-Domain	1,223	Text, Image	Snopes, Reuters	2
MOCHEG	Multi-Domain	15,601	Text, Image	Politifact, Snopes	3
Fakeddit	Multi-Domain	1,063,106	Text, Image	Reddit	2,3,6
FIN-FACT (ours)	Finance	3,369	Text, Image	Politifact, Snopes & FactCheck	3

a wide range of financial information. Additionally, it includes expert fact-checker comments, enabling comprehensive explanations by models.

Explanation Generation. Explanation generation plays a pivotal role in facilitating human comprehension of claim credibility. It involves leveraging external knowledge graphs to create semantic traces originating from the claim itself (Gad-Elrab et al., 2019; Li et al., 2020; Sarrouti et al., 2021). These semantic traces function as explanations that substantiate the veracity of claims. This approach offers valuable insights into the rationale behind the model’s decision-making, thereby fostering trust. Moreover, the process of explanation generation relies on drawing evidence from diverse sources (Atanasova et al., 2020; Hanselowski et al., 2019; Fan et al., 2020) to validate claims. However, this evidence is frequently comprised of isolated sentences extracted from extensive collections of documents, which can make it challenging for humans to interpret the broader context. To effectively generate explanations, a high-quality dataset annotated by humans is essential. This paper addresses the need for such a dataset.

3 The FIN-FACT Dataset

The FIN-FACT dataset presents a diverse array of labels that enhance the depth of analysis when evaluating financial claims. These labels contribute a multifaceted perspective to the fact-checking process, augmenting its analytical capabilities.

The dataset comprises essential attributes, including the *Claim* and *Author* labels, which respectively represent the core assertion and its source. Temporal context is introduced via the *Posted Date* attribute, while claim summaries are provided by the *Sci-digest* label. To further contextualize claims, *Justification* offers insights into their accuracy, *Evidence* presents supporting information linked through *Evidence link*, and *Image link* to address the visual dimension, and the *Issues* label highlights complexities within claims. Ultimately, the *Claim Label* categorizes claims as True, False, or NEI (Not Enough Information).

By amalgamating these labels, the dataset establishes a comprehensive and multidimensional resource. This resource accommodates textual, temporal, evidentiary, and visual components, all of which are imperative for a thorough evaluation of claims during the fact-checking process.

3.1 Data Collection and Preprocessing

PolitiFact¹, Snopes² and FactCheck³ are prominent online platforms dedicated to countering the spread of false information. These platforms engage professional fact-checkers to meticulously analyze and verify individual claims, subsequently producing articles that offer their conclusions supported by relevant evidence. In our study, we leveraged these platforms as our primary sources of data.

¹<http://politifact.com/>

²<http://snopes.com/>

³<http://factcheck.org>

To elaborate, we devised a systematic process to gather essential information from PolitiFact, Snopes and FactCheck websites. This encompassed the extraction of text-based claims and the assignment of corresponding truthfulness labels. Moreover, we retrieved both textual and visual evidence, along with their associated links, which contributed substantially to the assessment of claim accuracy. Through advanced keyword-based filtering mechanisms, we identified and isolated claims and articles that specifically pertained to financial domain. This process involved the careful selection of terms and phrases related to various financial sectors, including investment, taxation, and corporate finance. Furthermore, we ensured that the extracted claims were diverse in nature, encompassing a wide array of financial topics such as market trends, economic policies, financial disclosures, and fiscal regulations. By emphasizing this specialized filtration approach, we aimed to create a dataset that distinctly focuses on the intricacies of fact-checking within the financial landscape.

It's noteworthy that the initial claims were collected by journalists affiliated with these platforms. These claims originated from diverse sources, including online speeches, public statements, news articles, and social media posts. Importantly, the fact-checkers from these platforms played a pivotal role by providing truthfulness labels, pertinent evidence, references to corroborating sources, and the articles delivering their final verdict. This comprehensive approach ensured the thorough and reliable collection of data, reinforcing the credibility of our assessment of claim accuracy.

Following the data collection, we embarked on a rigorous data cleaning process to ensure the quality and reliability of our dataset. This process was crucial as it eliminated potential noise or inconsistencies in the data that could affect subsequent analysis. We removed any extra spaces within and between the claims to standardize the text data and facilitate easier text processing in later stages. Emojis were eliminated from the text. While emojis can sometimes convey meaningful information, their interpretation can be highly subjective and context-dependent. To maintain the objectivity of our analysis, we decided to exclude them from our dataset. Double quotations were removed from the text. These can often create confusion during text processing, especially when they are used within sentences. By removing them, we ensured

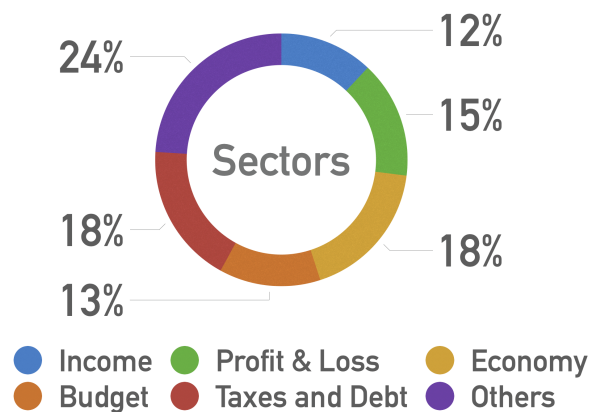


Figure 2: Diverse sectors within the FIN-FACT dataset.

that each claim is treated as a single continuous string of text, thereby simplifying the subsequent text analysis. Additional cleaning steps tailored to the specific characteristics of our dataset were conducted. These included the removal of HTML tags, correction of misspelled words, and standardization of text casing. Through these meticulous data cleaning procedures, we were able to refine our dataset and prepare it for the next stages of our research. This comprehensive approach to data cleaning not only enhanced the quality of our dataset but also reinforced the credibility of our subsequent analysis and findings.

3.2 Dataset Statistics

The FIN-FACT dataset is an encompassing compilation of claims within the financial domain, spanning diverse sectors such as Economy, Budget, Income, Taxes, and Debt, as visualized in Figure 2 and Figure 6 shows the most commonly occurring financial terms in the dataset. This dataset has been specifically constructed to ensure quality and relevance, comprising a total of 3,369 claims, curated to encapsulate the intricacies inherent in financial discourse.

In the FIN-FACT dataset, claims are categorized into three labels: True, False, and NEI (Not Enough Information) representing the veracity of each claim in the financial domain. The dataset contains 1,262 True claims that are verified as accurate, 1,462 False claims that have been proven inaccurate through fact-checking procedures, and 645 NEI instances where there is insufficient evidence to make a determination. With its comprehensive span across a variety of claims, diverse sectors, and an equitable distribution of labels, the FIN-FACT dataset serves as a new resource for research,

assessment, and progression of fact-checking models in the domain of finance.

4 Evaluation and Analysis

4.1 RQ1: Multi modal performance

In light of the rapid progress in Multimodal Language Models (MLLMs), there is a growing need for systematic evaluations of their capabilities across diverse tasks. A multitude of benchmark assessments has been developed to gauge the performance of MLLMs. Additionally, MLLMs are being leveraged for conducting multimodal fact-checking (Geng et al., 2024b), highlighting their versatility and applicability in this domain. We adopted a similar systematic evaluation approach to test the FIN-FACT dataset. We simultaneously obtained predictions, explanations, and confidence levels from MLLMs using the prompt below.

Is it true that {CLAIM}? True or False or NEI(Not Enough Information)? Use the following format to provide your answer:

Prediction: [True or False or NEI(Not Enough Information)]

Explanation: [put your evidence and step-by-step reasoning here]

Confidence Level: [please show the percentage]

We utilized OpenAI’s and Google’s APIs to gather responses from GPT-4V (gpt-4-1106-preview) and Gemini Vision Pro. Additionally, we conducted experiments with open-source multimodal language models (MLLMs), including LLaVA-1.5-13b (Liu et al., 2023) and InstructBLIP (Flan-T5-XXL) (Dai et al., 2023). All models were configured with default parameters, with the exception of max_tokens=500 for GPT-4V. These selected models are recognized for their representativeness and competitiveness within a multimodal context. Figure 3 provides a demonstration of prediction and explanation generation on the FIN-FACT dataset.

In our study, we have devised a classification framework based on the model’s response components—prediction, explanation, and confidence score. Despite having three primary labels (True, False, and NEI), the model occasionally produced alternative responses categorized as uncertain or deny. Uncertain responses explicitly convey the

model’s uncertainty using phrases like ‘inconclusive’, ‘unable to verify’, or ‘cannot determine’. The model denies if the claim contains information or elements that violate the specified input rules or criteria. Responses falling outside the primary labels are subsequently categorized based on the model’s prediction into True, False, or NEI. This classification framework aims to analyze the distribution of True, False, and NEI outcomes, assess the model’s compliance with instructions through Deny responses, and evaluate its level of uncertainty awareness indicated by the frequency of Uncertain declarations.

Upon comparing the data in Table 3, it is clear that images have a substantial impact on the predictions of both GPT-4V and Gemini Vision Pro, resulting in varying degrees of decreased accuracy. Concurrently, there is a noticeable increase in instances where the model opts not to respond.

Calibration refers to the alignment between a model’s predicted probabilities or confidence levels and the actual outcomes, demonstrating the model’s awareness and accuracy in estimating confidence (Geng et al., 2024a). In this assessment, we leverage verbalized confidence, which is readily accessible.

The calibration curves indicate a positive relationship between confidence levels and accuracy for both models. GPT-4V and Gemini Vision Pro exhibit calibration curves closely resembling the ideal curve (depicted by the dashed line), indicating well-calibrated confidence scores. Conversely, the calibration curve for LLaVA(13b) suggests a tendency towards overconfidence. Figure 4 shows the confidence score plot and the calibration curve.

4.2 RQ2: NLI performance

In the series of experiments, our focus extended to evaluating the performance of Natural Language Inference (NLI) models specifically in fact-checking tasks using text-only inputs. This evaluation aimed to assess how well NLI models could discern the veracity of financial statements and claims without relying on image inputs. We employed a range of well-known NLI models including GPT-2 (Radford et al., 2019), BART-Large (Lewis et al., 2020), RoBERTa-Large (Liu et al., 2019), and ELECTRA (Clark et al., 2020), as well as advanced Large Language Models (LLMs) like GPT-4, Gemini Pro, Claude-Opus, and Mixtral 8x7B (Jiang et al., 2024). Through evaluations using the FIN-FACT dataset,



Figure 3: Example demonstration of GPT-4V for prediction and explanation generation of real-world claims.

Model Type	Model	Precision	Recall	F1-Score	Accuracy	Modality
LVLM	Gemini Vision Pro	0.59 _{0.04}	0.57 _{0.03}	0.59 _{0.03}	0.58 _{0.06}	Text, Image
	GPT4-Vision	0.51 _{0.08}	0.48 _{0.06}	0.49 _{0.04}	0.53 _{0.08}	
	LLaVA(13b)	0.44 _{0.03}	0.40 _{0.01}	0.45 _{0.02}	0.50 _{0.05}	
	InstructBLIP	0.31 _{0.02}	0.28 _{0.03}	0.30 _{0.03}	0.37 _{0.05}	
LVLM	Gemini Vision Pro	0.51 _{0.03}	0.49 _{0.03}	0.50 _{0.04}	0.51 _{0.04}	Text
	GPT4-Vision	0.48 _{0.05}	0.44 _{0.06}	0.45 _{0.07}	0.50 _{0.04}	
	LLaVA(13b)	0.40 _{0.06}	0.38 _{0.08}	0.42 _{0.04}	0.43 _{0.07}	
	InstructBLIP	0.28 _{0.03}	0.25 _{0.05}	0.28 _{0.02}	0.33 _{0.03}	
NLI	GPT-2	0.347 _{0.07}	0.337 _{0.04}	0.312 _{0.05}	0.430 _{0.03}	Text
	BART-Large	0.376 _{0.04}	0.377 _{0.02}	0.344 _{0.07}	0.346 _{0.04}	
	RoBERTa-Large	0.352 _{0.01}	0.292 _{0.04}	0.255 _{0.04}	0.333 _{0.01}	
	ELECTRA	0.319 _{0.08}	0.300 _{0.07}	0.286 _{0.03}	0.297 _{0.05}	
LLM	GPT4	0.76 _{0.07}	0.79 _{0.04}	0.76 _{0.03}	0.78 _{0.05}	Text
	Claude3-Opus	0.62 _{0.04}	0.65 _{0.04}	0.61 _{0.04}	0.64 _{0.07}	
	Gemini-Pro	0.45 _{0.03}	0.42 _{0.05}	0.44 _{0.03}	0.47 _{0.06}	
	Mistral(8x7b)	0.42 _{0.09}	0.41 _{0.08}	0.48 _{0.08}	0.47 _{0.10}	

Table 3: Performance of various models with and without image input.

we aimed to comprehensively analyze these models' ability to provide accurate fact-checking outcomes using textual information alone.

The results of these fact-checking efforts have yielded interesting insights: Gemini Pro showcased 0.61, Claude3-Opus demonstrated 0.44, Mistral (8x7B) achieved F1-score of 0.48, and GPT-4 emerged as the leader with an F1-score of 0.76 as shown in Table 3. These findings underscore the intricate challenges posed by financial fact-checking, with models displaying varying degrees of perfor-

mance within this domain. Figure 7 illustrate the confusion matrices of the state-of-the-art NLI models.

Model	ROUGE-1	ROUGE-2	ROUGE-3	GLUE
GPT4-Vision	0.91	0.83	0.60	0.219
Gemini Vision Pro	0.87	0.73	0.48	0.110
BART-Large	0.81	0.61	0.40	0.066
XL-SUM	0.79	0.54	0.23	0.013
PEGASUS	0.68	0.41	0.16	0.009

Table 4: Performance scores of FIN-FACT on explanation generation models.

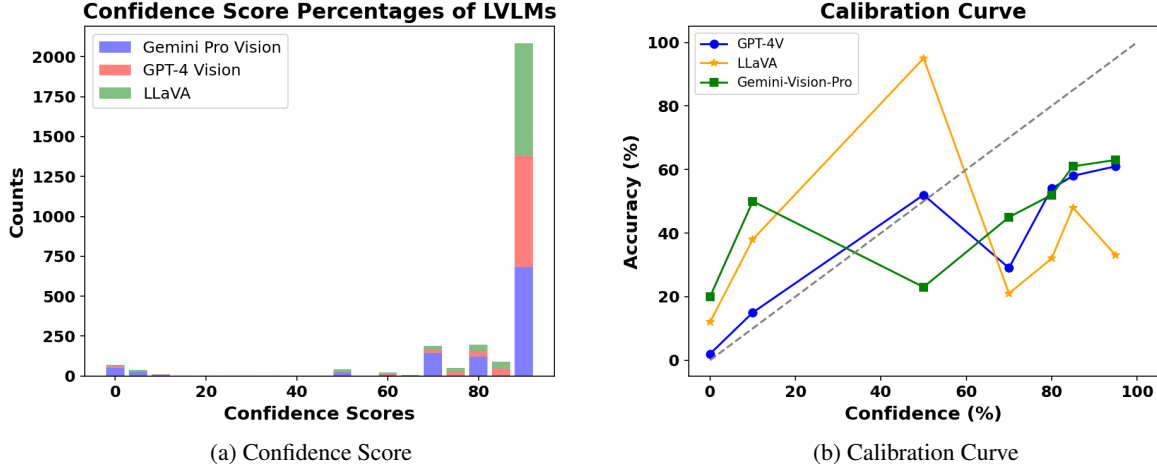


Figure 4: Confidence score and it’s calibration

4.3 RQ3: Explanation Generation

The final phase of our experimentation focused on generating explanations for the claims. For each claim in the dataset, we utilized large vision-language models (LVLMs) such as GPT-4 Vision, Gemini Vision Pro, along with language models like BART-Large (Lewis et al., 2020), XL-SUM (Hasan et al., 2021), and PEGASUS (Zhang et al., 2019) to generate detailed explanations. These insights highlighted the key factors influencing the determination of claim accuracy. However, InstructBLIP, despite its capabilities in other aspects, has limitations in providing explanations and confidence scores. These explanations were obtained using the justifications provided to the claim. To quantitatively evaluate the quality of these explanations, we leveraged the GLUE and ROUGE metrics as shown in Table 4. The Evidence label in the dataset served as the ground truth, enabling us to assess the alignment between the generated explanations and the human-provided justifications. We illustrate an example of explanation generation in Figure 5 and Table 6.

5 Limitations

We acknowledge certain limitations inherent in our study, especially the reliance solely on labels sourced from a fact-checking portal such as PolitiFact. Valid concerns have been raised regarding the handling of complex propositions and unverifiable claims. Moreover, we recognize the inherent constraints of Natural Language Processing (NLP) when addressing and refuting real-world misinformation, particularly during periods of uncertainty

	Claim	Label: FALSE
	Because of Obamacare, Medicare is going broke.	
Gold explanation	As U.S. House Speaker Paul Ryan discussed the repeal and replacement of the Affordable Care Act, a top priority of President-elect Donald Trump, he made a claim about the financial health of Medicare...	
BART explanation	U.S. House Speaker Paul Ryan made a claim about the financial health of Medicare. Ryan: Because of Obamacare, Medicare is going broke. Ryan's claim flies in the face of evidence that it actually helps shore up Medicare...	

Figure 5: Example of model-generated explanation as compared to the gold standard from FIN-FACT dataset.

where counter-evidence may not be readily available. It is acknowledged that current NLP methods may not match the effectiveness of professional fact-checkers in debunking misinformation.

While our study highlights the performance of the FIN-FACT dataset using large language models (LLMs) and large vision-language models (LVLMs), it is essential to address several limitations, particularly concerning the phenomenon of hallucination.

One significant challenge observed in our experiments is the tendency of LLMs and LVLMs to exhibit hallucinatory behaviors in their generated outputs. These models, trained on vast amounts of data, sometimes produce responses that seem contextually plausible but lack grounding in the

input information. This risk of hallucination becomes more pronounced in multimodal datasets, where information from multiple modalities (text, images) must be integrated coherently. LVLMS may generate explanations for images that include details not present in the visual content, indicating an over-reliance on textual patterns rather than visual cues.

Another challenge arose when the watermark ‘SCAM’ was present on the images, as illustrated in the first claim of Figure 3. In some instances, the model’s explanations referenced the watermark. For example, the model’s explanation might mention, "Additionally, the advertisement is labeled with a ‘SCAM’ warning, suggesting that it has been identified as a fraudulent scheme."

6 Conclusion and Future Work

The advent of FIN-FACT marks a significant stride forward in the battle against misinformation within the financial sphere. Through its incorporation of expert annotations, extensive claims data, and the promise of in-depth explanatory insights, FIN-FACT equips fact-checking systems with the tools needed to attain heightened levels of precision and transparency. Its interdisciplinary framework effectively navigates the intricacies of financial language and the ever-evolving contextual complexities, establishing itself as a sturdy cornerstone for the enhancement of more effective and dependable fact-checking processes.

In the pursuit of advancing fact-checking capabilities and considering the increasing prevalence of misinformation dissemination through various multimedia channels, our future endeavors will focus on enhancing FIN-FACT to encompass a broader range of multimedia data, including video content, to foster a more comprehensive understanding of complex financial claims. Additionally, our future research will prioritize the development of robust methodologies to identify, analyze, and counteract any biases introduced through the visual elements. By addressing these critical concerns, we aim to enhance the reliability of fact-checking processes in the financial domain.

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. [Generating fact checking explanations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, pages 7352–7364, Online. Association for Computational Linguistics.

Christoph E Boehm and T. Niklas Kroner. 2023. [The us, economic news, and the global financial cycle](#). Working Paper 30994, National Bureau of Economic Research.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. [Electra: Pre-training text encoders as discriminators rather than generators](#).

Jonathan Clarke, Hailiang Chen, Ding Du, and Yu Jeffrey Hu. Forthcoming. Fake news, investor attention, and market reaction. *Information Systems Research*. 45 Pages Posted: 1 Aug 2018 Last revised: 18 Dec 2019.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. [Instructblip: Towards general-purpose vision-language models with instruction tuning](#).

Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. 2020. Generating fact checking briefs. *arXiv preprint arXiv:2011.05448*.

Mohamed H Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. 2019. Exfakt: A framework for explaining facts over knowledge graphs and text. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 87–95.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koepl, Preslav Nakov, and Iryna Gurevych. 2024a. [A survey of confidence estimation and calibration in large language models](#).

Jiahui Geng, Yova Kementchedjheva, Preslav Nakov, and Iryna Gurevych. 2024b. [Multimodal large language models to support real-world fact-checking](#).

Ashim Gupta and Vivek Srikumar. 2021. X-FACT: A New Benchmark Dataset for Multilingual Fact Checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. *arXiv preprint arXiv:1911.01214*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. [XLsum: Large-scale multilingual abstractive summarization for 44 languages](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

- Md Zobaer Hossain, Md Ashraful Rahman, Md Sai-ful Islam, and Sudipta Kar. 2020. Banfakenews: A dataset for detecting fake news in bangla. *arXiv preprint arXiv:2004.08789*.
- Andreas Hüsser and Werner Wirth. 2014. Do investors show an attentional bias toward past performance? an eye-tracking experiment on visual attention to mutual fund disclosures in simplified fund prospectuses. *Journal of Financial Services Marketing*, 19:169–185.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Shimon Kogan, Tobias J Moskowicz, and Marina Niessner. 2019. Fake news: Evidence from financial markets. Available at SSRN, 3237763.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. *arXiv preprint arXiv:2011.04088*.
- Betty Liu and Austin Moss. 2022. The role of accounting information in an era of fake news. Available at SSRN 4399543.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. [Improved baselines with visual instruction tuning](#).
- 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](#).
- Hamid Mansoor and Lane Harrison. 2018. Data visualization literacy and visualization biases: Cases for merging parallel threads. *Cognitive biases in visualizations*, pages 87–96.
- Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. 2022. Factify: A multi-modal fact verification dataset. In *Proceedings of the First Workshop on Multimodal Fact-Checking and Hate Speech Detection (DE-FACTIFY)*.
- Padmapriya Mohankumar, Ashraf Kamal, Vishal Kumar Singh, and Amrish Satish. 2023. Financial fake news detection via context-aware embedding and sequential representation using cross-joint networks. In *2023 15th International Conference on COMMunication Systems & NETWORKS (COMSNETS)*, pages 780–784. IEEE.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Aman Rangapur, Haoran Wang, and Kai Shu. 2023. [Investigating online financial misinformation and its consequences: A computational perspective](#).
- Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. [COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.
- Mourad Sarrouiti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. [Evidence-based fact-checking of health-related claims](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- 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.
- David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. [Fact or fiction: Verifying scientific claims](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.
- David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. [SciFact-open: Towards open-domain scientific claim verification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. [Pegasus: Pre-training with extracted gap-sentences for abstractive summarization](#).

Xiaohui Zhang, Qianzhou Du, and Zhongju Zhang. 2022. A theory-driven machine learning system for financial disinformation detection. *Production and Operations Management*, 31(8):3160–3179.

Xiaofan Zhi, Li Xue, Wengang Zhi, Ziyi Li, Bo Zhao, Yanzhen Wang, and Zhen Shen. 2021. Financial fake news detection with multi fact cnn-lstm model. In *2021 IEEE 4th International Conference on Electronics Technology (ICET)*, pages 1338–1341. IEEE.

A Appendix

A.1 Additional Details of Dataset

Data visualization serves as a crucial role for conveying insights and information, particularly in the realm of financial misinformation fact-checking. Visual representations of data can illuminate complex patterns and trends, aiding in the identification and debunking of false claims. However, it is imperative to acknowledge the dual nature of data visualization, especially within the context of misinformation. While visualizations are instrumental in clarifying and contextualizing data, they can also be manipulated to propagate misleading narratives. In the domain of financial misinformation, the stakes are particularly high, as distorted visualizations can significantly impact public perception and policy decisions.

The potency of data visualization lies in its ability to simplify complex data sets, making them accessible to diverse audiences. Yet, this very attribute renders visualizations susceptible to manipulation. Misleading techniques such as selective data omission, altered scales, or misleading labeling can skew interpretations and perpetuate misinformation. Moreover, sophisticated graphical representations can obscure underlying inaccuracies, making it challenging for viewers to discern fact from fiction.

To combat the misuse of data visualization in the propagation of financial misinformation, a critical approach is required. This involves not only verifying the accuracy of the underlying data but also scrutinizing the design and presentation of visualizations. Transparency in visualization practices, coupled with robust fact-checking methodologies, is essential in safeguarding against the dissemination of false narratives.

A.2 Additional Evaluation Details

In addition to the primary evaluation metrics discussed earlier, we believe in providing a comprehensive assessment of our models’ capabilities. To

achieve this, we delve deeper into the performance analysis, offering a more nuanced understanding of our models’ effectiveness. This comprehensive evaluation includes the utilization of confusion matrices and classification reports, specifically tailored for the Natural Language Inference (NLI) models that play a pivotal role in our fact-checking endeavors. We employ confusion matrices to scrutinize our NLI models’ performance at a finer level of detail.

Beyond confusion matrices, we present comprehensive classification reports for each of our NLI models. These reports go beyond mere accuracy metrics and provide in-depth assessments of the model’s performance. For every label (‘True,’ ‘False,’ and ‘NEI’), we furnish precision, F1-score, and recall values. Such detailed reports not only offer a holistic view of each model’s overall accuracy but also highlight its strengths and weaknesses in classifying claims within the intricate domain of finance. Table 5 illustrates the precision, recall, and F1-score of various NLI models for each label.

Table 5: Classification Metrics when performed on NLI Models.

Model	Precision	Recall	F1-Score	Label
GPT-2	0.49	0.71	0.58	True
	0.14	0.17	0.15	NEI
	0.41	0.13	0.30	False
BART-Large	0.44	0.37	0.40	True
	0.48	0.25	0.33	NEI
	0.21	0.50	0.30	False
RoBERTa-Large	0.42	0.39	0.40	True
	0.49	0.27	0.35	NEI
	0.20	0.43	0.27	False
ELECTRA	0.49	0.31	0.38	True
	0.12	0.50	0.19	NEI
	0.44	0.18	0.25	False

A.3 Evaluation Setup and Reproducibility

A.3.1 Hyperparameter Settings

In the process of evaluating the corpus with Natural Language Inference (NLI) models using the Transformers library from Hugging Face, we meticulously explored various hyperparameter settings to ascertain the most effective configuration. Among these settings, the `max_length` parameter of the tokenizer was subjected to thorough experimentation. We experimented with different values to strike a delicate balance between capturing crucial information and ensuring efficient processing. After rigorous testing, it became evident that a `max_length`

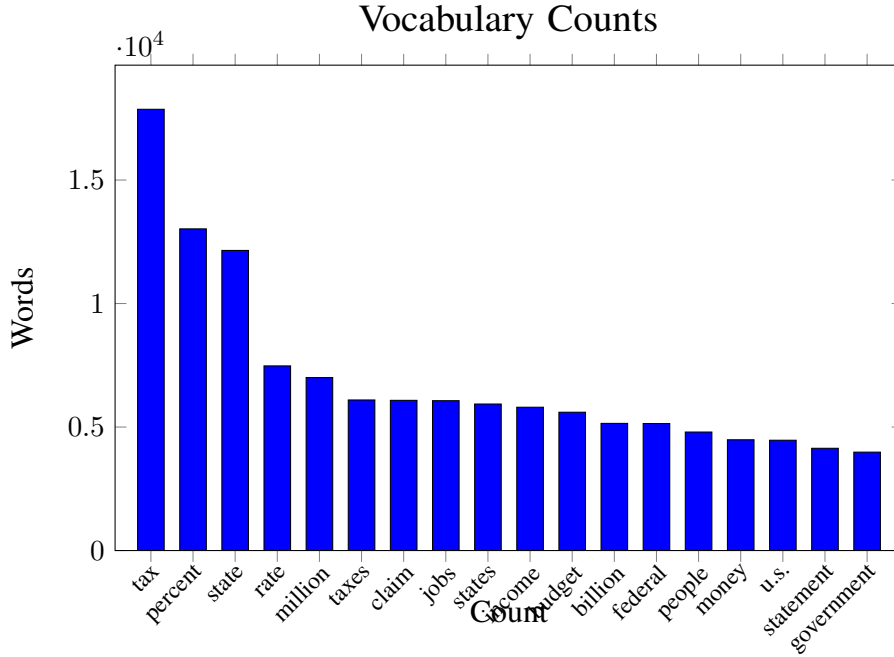


Figure 6: Frequency of vocabulary in FIN-FACT dataset.

value of 256 yielded the most promising results. This choice was the culmination of a systematic exploration, aiming to align our settings with industry best practices and to extract optimal performance from the models.

A.3.2 Computing Infrastructure

The conducted experiments were carried out leveraging the computational prowess of the NVIDIA RTX 3070 GPU with 16 GB VRAM, meticulously chosen for its robust processing capabilities. The overarching objective of these experiments was to precisely evaluate the performance of a diverse spectrum of models on the FIN-FACT dataset. The computational resources furnished by the NVIDIA RTX 3070 GPU served as a cornerstone, ensuring the dependable and efficient execution of the experimental procedures. The average evaluation time taken by NLI models was 45 minutes, while the explanation generation models took around 4 hours.

		Confusion Matrix		
		False	NEI	True
Actuals	False	366	676	419
	NEI	132	325	188
	True	268	522	473
		Predictions		

(a) Confusion matrix of BART on FIN-FACT.

		Confusion Matrix		
		False	NEI	True
Actuals	False	235	653	427
	NEI	66	219	155
	True	229	1009	569
		Predictions		

(b) Confusion matrix of ELECTRA on FIN-FACT.

		Confusion Matrix		
		False	NEI	True
Actuals	False	400	561	500
	NEI	186	277	182
	True	237	539	487
		Predictions		

(c) Confusion matrix of RoBERTa on FIN-FACT.

		Confusion Matrix		
		False	NEI	True
Actuals	False	175	149	991
	NEI	43	75	322
	True	211	314	1282
		Predictions		

(d) Confusion matrix of GPT-2 on FIN-FACT.

Figure 7: Confusion matrices of different NLI models on FIN-FACT.

Table 6: Examples illustrating explanation generation using BART-Large for various statements.

Title	Back in the Great Recession, when millions of jobs across America hung in the balance, Donald Trump said rescuing the auto industry didn't really matter very much. He said, and I quote again, Let it go.
Justification	Donald Trump didn't care about rescuing the auto industry during the 2008 recession, Hillary Clinton said at a rally in Michigan. Clinton told her audience at Wayne State University in Detroit that she supported the 2008 government bailout for the American auto industry...
Evidence	Trump's position on an auto bailout was inconsistent leading up to Dec. 19, 2008, when Bush laid out plans to go forward with the bailout. Trump said the government should help the auto companies, but he also regularly suggested the they could save themselves if they filed for...
Generated Fact	Hillary Clinton said at a rally in Michigan that she supported the 2008 government bailout for the American auto industry. But Trump, she said, doesn't support American workers, much less those in Michigan's Auto industry. Trump's public comments about the auto industry's demise were the opposite of what Clinton said.
Title	Video shows that George Soros is going bankrupt.
Justification	We found no news reports that he's heading into bankruptcy or that Musk leaked this information. A narrator in the Facebook posts video says that in a recent interview, Musk revealed a piece of information related to the existence of George Soros and his Open Society Foundation, which is on the verge of bankruptcy...
Evidence	IT HAPPENED, the July 18 post says. Elon Musk L.E.A.K.E.D GEORGE SOROS IS GOING B.A.N.K.R.U.P.T, JOE BIDEN S.C.A.RED! - BREAKING FOX NEWS BIDEN - JULY 19, 2023. This post was flagged as part of Meta's efforts to combat false news and misinformation on its News Feed...
Generated Fact	A recent Facebook post claims that billionaire philanthropist George Soros is in the red, but it provides no evidence that this is true. According to Forbes, Soros was worth \$6.7 billion as of July 20. This post was flagged as part of Meta's efforts to combat false news and misinformation on its News Feed.
Title	Because of Obamacare, Medicare is going broke.
Justification	What the ACA did on Medicare: Here's how Marc Goldwein, senior policy director at the Committee for a Responsible Federal Budget, lays it out: On one side of the ledger, Obamacare slowed the growth of Medicare spending by, among other things, reducing increases in payments made to nearly all health care providers.
Evidence	As U.S. House Speaker Paul Ryan discussed the repeal and replacement of the Affordable Care Act, a top priority of President-elect Donald Trump, he made a claim about the financial health of Medicare. "What people don't realize is, because of Obamacare, Medicare is going broke," the Wisconsin Republican told Fox News host Bret Baier on Nov. 10, 2016. The Affordable Care Act certainly has its detractors...
Generated Fact	U.S. House Speaker Paul Ryan made a claim about the financial health of Medicare. Ryan: Because of Obamacare, Medicare is going broke. Ryan's claim flies in the face of evidence that it actually helps shore up Medicare. On one side of the ledger, Obamacare slowed the growth of Medicare spending.