EXPLORING THE RELIABILITY OF SELF-EXPLANATION AND ITS RELATIONSHIP WITH CLASSIFICATION IN LANGUAGE MODEL-DRIVEN FINANCIAL ANALYSIS

Han Yuan & Li Zhang & Zheng Ma[✉]
American Express Global Decision Science
{Han.Yuan1, Li.Zhang1, Zheng.Ma2}@aexp.com

ABSTRACT

Language models (LMs) have exhibited exceptional versatility in reasoning and in-depth financial analysis through their proprietary information processing capabilities. Previous research focused on evaluating classification performance while often overlooking explainability or pre-conceived that refined explanation corresponds to higher classification accuracy. Using a public dataset in finance domain, we quantitatively evaluated self-explanations by LMs, focusing on their factuality and causality. We identified the statistically significant relationship between the accuracy of classifications and the factuality or causality of self-explanations. Our study built an empirical foundation for approximating classification confidence through self-explanations and for optimizing classification via proprietary reasoning.

1 INTRODUCTION

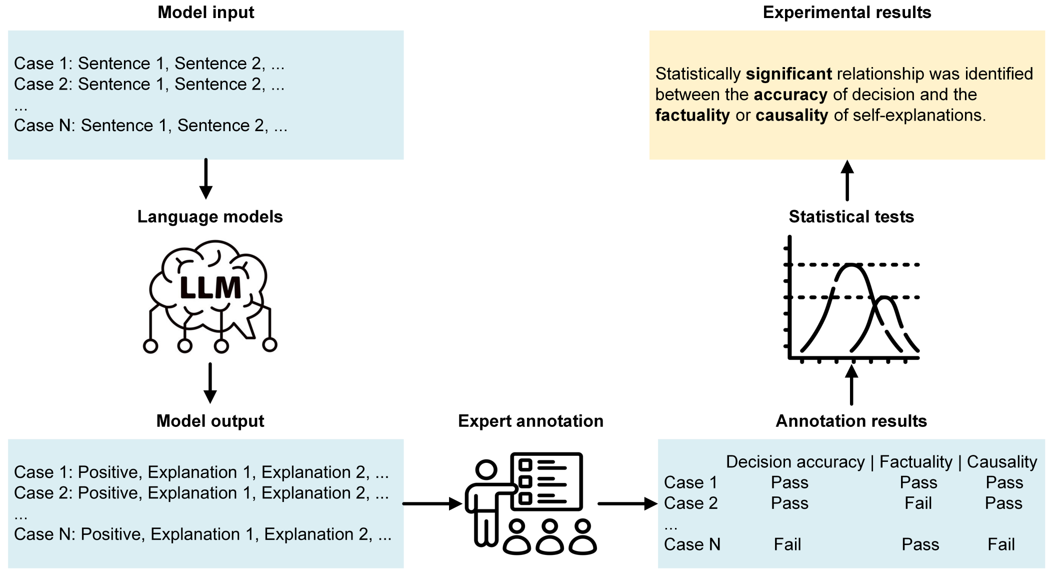
Recent advances in model architectures, computing hardware, and data resources have positioned language models (LMs) as versatile problem solvers in various domains, and previous studies have also highlighted the potential of LMs in financial classification (Lee et al., 2024; Kirtac & Germano, 2024; Arslan et al., 2021; Xie et al., 2024; Fatemi et al., 2025). In formal terms, classification by LMs differs from open-ended generation in that it involves structured generation constrained by a predefined set of options. For instance, in stock trading, classifications are typically limited to two discrete actions: long or short (Chuang & Yang, 2022; Koa et al., 2024; Bao et al., 2024). In such cases, LMs are required to provide a single, definitive response rather than ambiguous recommendations, such as suggesting that both options could be reasonable under certain conditions.

Previous research on financial classification has primarily focused on improving the main target: accuracy of classifications (Guo et al., 2023; Chen et al., 2024a; Li et al., 2023a). However, explanations of classifications also warrant attention. A classification without accompanying explanation can lead to severe consequences in finance, such as asset losses in the stock trading example. In addition, the absence of explanations makes it challenging for human experts to promptly assess the validity of classifications made by LMs, potentially undermining trust and utility.

Recent advancements such as DeepSeek (Liu et al., 2024; Lu et al., 2024; Liu et al., 2024) show the potential of self-explanations and reasoning to enhance model performance without explicit user instructions. While overall performance gains have been observed, a quantitative evaluation of the relationship between self-explanations and generation quality is essential to establish empirical evidence supporting the reliability of this approach. Some studies have explored the role of proprietary explanations and reasoning in enhancing generation by LMs (Zhao et al., 2023). Lampinen et al. (2022) examined the effect of providing a few in-context examples to LMs' prompts and concluded that explanations improved model performance in general domains. Furthermore, Ye & Durrett (2024) investigated the use of triplets comprising a question, classification, and explanation in few-shot examples, demonstrating LMs tend to generate nonfactual explanations when making wrong predictions.

[✉] Correspondence: Zheng Ma, Singapore Decision Science Center of Excellence, American Express, 1 Marina Boulevard, 018989, Singapore.

Our study, as shown in Figure 1 extends this line of research in three key ways. First, we focused on the financial domain, where classification errors can lead to substantial real-world consequences. Second, we investigated zero-shot classification scenarios, where explanations, referred to as self-explanations, were generated by LMs without the aid of in-context examples. Third, we conducted detailed annotations of self-explanations in these scenarios to quantify the relationship between the accuracy of classifications and the factuality or causality of self-explanations, statistically confirming the consistency of observations reported by Ye & Durrett (2024) within the financial domain.


Model input
Case 1: Sentence 1, Sentence 2, ...
Case 2: Sentence 1, Sentence 2, ...
...
Case N: Sentence 1, Sentence 2, ...
Language models
LLM
Model output
Case 1: Positive, Explanation 1, Explanation 2, ...
Case 2: Positive, Explanation 1, Explanation 2, ...
...
Case N: Positive, Explanation 1, Explanation 2, ...
Expert annotation
Annotation results
Statistical tests
Experimental results
Statistically significant relationship was identified between the accuracy of decision and the factuality or causality of self-explanations.

	Decision accuracy	Factuality	Causality
Case 1	Pass	Pass	Pass
Case 2	Pass	Fail	Pass
...			
Case N	Fail	Pass	Fail

Figure 1: Schematic plot of our experimental pipeline

Our experimental finding established an empirical foundation for advancing two research directions. First, it enables the development of a proxy confidence metric of the classification accuracy, such as the proportion of factual or causal inconsistencies in the explanations for a given case. In practical scenarios, such as classifying the hawkish or dovish stance in Federal Open Market Committee (FOMC) speeches, even experts struggle to make highly accurate classifications. Consequently, they cannot reliably assess whether the classifications made by LMs are trustworthy while the factuality and causality of explanations would be more easily to evaluate. With the proxy confidence value, users can make more informed choices regarding the adoption of LMs' classifications. The second area is to provide the empirical foundation for financial classification optimization based on LMs' explanations, which can be interpreted as proprietary reasoning underlying classifications. If no correlation exists between LMs' explanations and their classification accuracy, it becomes difficult to justify the premise that improving model explanations or reasoning would enhance financial classification (Shinn et al., 2024; Dou et al., 2024; Liu et al., 2024).

The remainder of this paper is structured as follows: Section 2.1 presents the public dataset information and relevant data processing. Section 2.2 details the LMs and other configurations used in this study. Sections 2.3 and 2.4 analyze the statistically significant relationship between the accuracy of classifications and the factuality or causality of self-explanations in zero-shot classification by LMs. Section 3 discusses the limitations of the present study and outlines the planned work informed by current findings. Section 4 concludes the study and provides recommendations for future research. Disclaimers are attached at the end and Appendix A shows the detailed classification performance.

2 Experiments

2.1 Data

We conducted the experiments using the publicly available German credit dataset (Hofmann, 1994), a widely recognized benchmark in financial natural language processing (NLP). To reflect the actual classification capability of LMs, we refined the dataset to increase its signal-to-noise ratio and make it better aligned with LMs' training context. All processing steps were entirely ad hoc and did not involve any analytical operations related to the label, ensuring that performance was not influenced by information leakage. For reproduction, the processed dataset in textual format is released and Appendix A demonstrates the effectiveness of data processing. Also, all experiments were conducted on all minority cases paired with an equal number of majority cases to ensure that the analyses were not adversely affected by the data imbalance (Yuan et al., 2022).

2.2 Language models

We utilized three general-purpose LMs: Meta's Llama-3.2-3B (Touvron et al., 2023), Microsoft's Phi-3.5-3.8B (Abdin et al., 2024), and Google's Gemma-2-2B Mesnard et al. (2024). All LMs were utilized in their instruction-tuned versions, with computations in half-precision due to hardware limitations. The code is publicly available for reproduction.

2.3 Factuality

In financial analysis, beyond answers, it is crucial to provide explanations to meet regulatory requirements and support classification recalibration by financial service providers, thereby mitigating potential adverse impacts. A notable advantage of LMs is their ability to generate user-friendly explanations in natural language. However, the quality of these self-explanations should be rigorously evaluated, as self-explanations containing factual inaccuracies, a prevalent issue of LMs, are of little practical value (Ji et al., 2023). Moreover, if statistically significant relationships exist between the accuracy of classifications and the presence of fabricated information in self-explanations, factuality could serve as a proxy for assessing the confidence of classifications made by LMs.

The classifications and self-explanations generated by the three LMs were analyzed across top 100 cases, comprising 50 positive and 50 negative cases. Detailed annotations have been released. We quantified the prevalence of factuality issues as the ratio of cases in which the explanations contained any factual inaccuracies. An explanation was classified as having a factuality issue if even a single sentence within it was factually incorrect. Further, we conducted a Chi-squared test to evaluate the independence of factuality issues from the accuracy of model classifications. The null hypothesis posited no relationship between the factuality of the self-explanations and the accuracy of the classifications. The P value is lower than 0.05, suggesting statistically significant dependence between the occurrence of factuality and the accuracy of classification at a confidence level of 95%.

Table 1 presents an analysis of the relationship between factuality and the accuracy of LMs' classifications. Chi-squared tests revealed statistically significant associations between these factors (P $\leq$ 0.05). We also quantified the prevalence of factuality issues in self-explanations generated by different LMs for both positive and negative cases, observing that Llama-3.2-3B exhibited fewer factuality issues compared to the other two LMs.

Table 1: Impact analysis of factuality of self-explanations on the accuracy of classifications

Language model	Case type	Issue prevalence	Chi-square statistic	P value
Llama-3.2-3B	Positive	0.16	23.28	3.53e-05
	Negative	0.20	72.40	1.30e-15
Phi-3.5-3.8B	Positive	0.26	67.28	1.63e-14
	Negative	0.22	81.36	1.57e-17
Gemma-2-2B	Positive	0.32	57.68	1.84e-12
	Negative	0.16	96.24	1.00e-20

https://github.com/Han-Yuan-Med/Language-Models-for-Finance/tree/main/Dataset

https://github.com/Han-Yuan-Med/Language-Models-for-Finance/tree/main/Code

https://github.com/Han-Yuan-Med/Language-Models-for-Finance/tree/main/Annotation

2.4 Causality

After the factuality check, another critical issue with self-explanations is causality, also referred to as logical inconsistency (Ye & Durrett, 2024). Specifically, the reasoning within a self-explanation may be contradictory, such as when negative attributes are described as contributing to a positive classification, or vice versa. Similar to factuality, causality is essential for assessing the usability of LMs' self-explanations and serves as a potential confidence measure for the reliability of classifications made by LMs. The same Chi-squared test was computed.

The statistically significant relationship was observed between the causality of self-explanations and the accuracy of classifications across diverse LMs in Table 2. Also, compared to factuality issues, causality issues were more prevalent in the self-explanations generated by LMs. Among the evaluated models, Gemma-2-2B demonstrated fewer causality issues than the other two LMs. Besides, in most scenarios, while both factuality and causality exhibited statistically significant relationship with classification accuracy, factuality demonstrated a stronger correlation. This finding suggested that factuality, compared with causality, could serve as a better confidence proxy for classification.

Table 2: Impact analysis of causality on the accuracy of model classifications

Language model	Case type	Issue prevalence	Chi-square statistic	P value
Llama-3.2-3B	Positive	0.80	22.00	6.52e-05
	Negative	0.82	72.72	1.12e-15
Phi-3.5-3.8B	Positive	0.52	46.68	4.48e-10
	Negative	0.68	62.96	1.37e-13
Gemma-2-2B	Positive	0.52	46.48	4.48e-10
	Negative	0.52	50.16	7.29e-11

3 Discussion

This study highlights the statistically significant relationship between the accuracy of classifications and factuality or causality of self-explanations. Due to computational constraints, we did not extend our experiments to LMs such as Llama-3.2-90B. Future research should explore these large-scale models, focusing on comparisons with our current findings in terms of factuality and causality. This would help determine whether the increase in susceptibility to hallucination with larger model parameters also hold in financial applications (Rawte et al., 2023; Li et al., 2023b). Also, our experiments were limited to a single task. Future study should evaluate a broader range of tasks to verify the generalizability of our findings.

Building on extensive validation of the statistically significant relationship between the classification accuracy and the factuality or causality of self-explanations across diverse financial NLP tasks, we scheduled two further studies: (1) to fine-tune an automated detector for accurately identifying factuality or causality issues within self-explanations and test its utility as a confidence proxy for LMs' classifications, and (2) to benchmark existing reasoning-enhancement strategies (Chen et al., 2024b; Hao et al., 2023) based on automatically identified errors in optimizing financial applications.

4 Conclusion

In this study, we investigated the use of LMs for financial classification. Beyond final classifications, LMs demonstrated the ability to provide human-interpretable explanations. While challenges related to factuality and causality exist in self-explanations, recent advancements in LMs have substantially mitigated the factuality issue. With the continued evolution of reasoning capabilities, LMs hold promise for delivering highly accurate and self-explainable classifications in finance.

Disclaimer

This paper is intended solely for informational purposes and is not a product of or intended to constitute any business practice of American Express. The opinions, findings and conclusions of this paper are those of the authors alone and do not reflect the views of American Express.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

Yusuf Arslan, Kevin Allix, Lisa Veiber, Cedric Lothritz, Tegawendé F Bissyandé, Jacques Klein, et al. A comparison of pre-trained language models for multi-class text classification in the financial domain. In *Proceedings of the Web Conference*, pp. 260–268, 2021.

Wuzhida Bao, Yuting Cao, Yin Yang, Hangjun Che, Junjian Huang, and Shiping Wen. Data-driven stock forecasting models based on neural networks: A review. *Information Fusion*, pp. 102616, 2024.

Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. FinTral: A family of GPT-4 level multimodal financial large language models. In *Findings of the Association for Computational Linguistics*, pp. 13064–13087, 2024.

Jian Chen, Peilin Zhou, Yining Hua, Loh Xin, Kehui Chen, Ziyuan Li, et al. FinTextQA: A dataset for long-form financial question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 6025–6047, 2024a.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Measuring and improving chain-of-thought reasoning in vision-language models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 192–210, 2024b.

Chengyu Chuang and Yi Yang. Buy tesla, sell ford: Assessing implicit stock market preference in pre-trained language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 100–105, 2022.

Zi-Yi Dou, Cheng-Fu Yang, Xueqing Wu, Kai-Wei Chang, and Nanyun Peng. Re-ReST: Reflection-reinforced self-training for language agents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 15394–15411, 2024.

Sorouralsadat Fatemi, Yuheng Hu, and Maryam Mousavi. A comparative analysis of instruction fine-tuning large language models for financial text classification. *Management Information Systems*, 2025.

Yue Guo, Zian Xu, and Yi Yang. Is ChatGPT a financial expert? evaluating language models on financial natural language processing. In *Findings of the Association for Computational Linguistics*, pp. 815–821, 2023.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, et al. Reasoning with language model is planning with world model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 8154–8173, 2023.

Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Kemal Kirtac and Guido Germano. Sentiment trading with large language models. *Finance Research Letters*, 62:105227, 2024.

Kelvin JL Koa, Yunshan Ma, Ritchie Ng, and Tat-Seng Chua. Learning to generate explainable stock predictions using self-reflective large language models. In *Proceedings of the ACM Web Conference*, pp. 4304–4315, 2024.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, et al. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics*, pp. 537–563, 2022.

David Kuo Chuen Lee, Chong Guan, Yinghui Yu, and Qinxu Ding. A comprehensive review of generative ai in finance. *FinTech*, 3(3):460–478, 2024.

Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, et al. Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? a study on several typical tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 408–422, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, et al. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023b.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv*, 2024.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv*, 2024.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, et al. Gemma: Open models based on gemini research and technology, 2024.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, et al. LILA: A unified benchmark for mathematical reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 5807–5832, 2022.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, et al. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 2541–2573, 2023.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, et al. Llama: Open and efficient foundation language models, 2023.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, et al. Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pp. 33469–33484, 2023.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, et al. Finben: A holistic financial benchmark for large language models. *Proceedings of the Advances in Neural Information Processing Systems*, pp. 95716–95743, 2024.

Xi Ye and Greg Durrett. The unreliability of explanations in few-shot prompting for textual reasoning. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2024.

Han Yuan, Feng Xie, Marcus Eng Hock Ong, Yilin Ning, Marcel Lucas Chee, et al. Autoscore-imbalance: An interpretable machine learning tool for development of clinical scores with rare events data. *Journal of Biomedical Informatics*, 129:104072, 2022.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the ACM on Conference on Information and Knowledge Management*, pp. 1569–1578, 2017.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 5823–5840, 2023.

A Appendix

The prior research Xie et al. (2023); Bhatia et al. (2024) on this dataset overlooked the critical role of data processing in enhancing the signal-to-noise ratio and revealing the true capabilities of LMs. Specifically, the original dataset includes outdated information and pre-existing bias (Zehlike et al., 2017) that pose challenges for LMs. For instance, certain features are denominated in Deutsche Marks, a currency that has been obsolete for over two decades. Also, LMs often exhibit limited sensitivity to numeric reasoning (Mishra et al., 2022). To address these issues, we excluded features misaligned with contemporary societal contexts where LMs were developed and converted numeric features into percentile representations through binarization.

For evaluation metrics, we adopted standard evaluation metrics of accuracy and F1 score. In addition, financial classification prioritizes weighted costs, emphasizing the greater consequence of false positive to false negative. As specified in the original dataset documentation Hofmann (1994), the cost associated with a false negative is quantified as 5, while that of a false positive is 1. A lower cost indicates superior performance.

For LMs selection, we reported the results of three LMs in the main text along with two financial domain-specific LMs, FinMA-7B (Xie et al., 2023) and FinTral-7B (Bhatia et al., 2024), which have comparable scales to the general-purpose models, based on their original publications.

Table 3 shows a consistent improvement in weighted cost across all three LMs when using the processed dataset. For accuracy and F1 score, a substantial enhancement was achieved with Llama-3.2-3B, while performance remained comparable for the other two LMs. Notably, in comparison to the instruction-tuned FinMA-7B and FinTral-7B models on the original German dataset, the zero-shot Llama-3.2-3B demonstrated superior performance, highlighting the effectiveness of data processing and ensuring the analytical quality of LMs' self-explanations.

Table 3: LMs performance on original and processed data

Model input	Language model	Weighted cost ↓	Accuracy ↑	F1 score ↑
Original data	Llama-3.2-3B	1185	0.53	0.34
	Phi-3.5-3.8B	304	0.51	0.67
	Gemma-2-2B	322	0.51	0.67
	FinMA-7B	-	-	0.17
	FinTral-7B	-	0.61	-
Processed data	Llama-3.2-3B	306	0.68	0.74
	Phi-3.5-3.8B	295	0.51	0.67
	Gemma-2-2B	296	0.51	0.67