

Numbers Matter: Leveraging Expert-Designed Hints to Improve Financial Sentiment Analysis

Chung-Chi Chen,¹ Hiroya Takamura,¹ Ichiro Kobayashi,² Yusuke Miyao³

¹ Artificial Intelligence Research Center, AIST, Japan

² Ochanomizu University, Japan

³ University of Tokyo, Japan

c.c.chen@acm.org, takamura.hiroya@aist.go.jp, koba@is.ocha.ac.jp, yusuke@is.s.u-tokyo.ac.jp

Abstract

Financial sentiment analysis requires understanding nuanced cues embedded in financial texts, especially numerical information. While large language models (LLMs) have shown strong performance in zero-shot tasks, it remains unclear whether they naturally attend to domain-specific indicators such as monetary values or percentages. In this study, we investigate whether expert-designed hints that highlight the importance of numbers can improve the zero-shot sentiment classification performance of LLMs. Using the Fin-SoMe dataset and evaluating four major LLMs (PaLM 2, Gemini Pro, GPT-3.5, and GPT-4), we compare standard zero-shot prompts, chain-of-thought (CoT) reasoning, and CoT with numerical focus hints. Our results show that models significantly benefit from expert-guided hints, especially on tweets containing numerical data or requiring deeper perspective-taking. We further analyze performance across different number categories, revealing that monetary and temporal expressions benefit most from explicit prompting. These findings suggest that even advanced LLMs can benefit from targeted guidance in domain-specific tasks.

Introduction

Sentiment analysis plays a crucial role in understanding opinions expressed in financial texts and has long been an active area in Natural Language Processing (NLP) (Baker and Wurgler 2007; Liu 2015; Xu and Cohen 2018; Chen, Huang, and Chen 2018). A key aspect that distinguishes financial language is the frequent use of numerical information—such as prices, percentages, and quantities—which many domain experts consider essential for interpreting sentiment in this context. Prior research in financial NLP has explored how numbers can be extracted (Chen et al. 2018; Yang et al. 2022; Shi et al. 2023) and reasoned (Zhu et al. 2021; Chen et al. 2021; Nan et al. 2022) about in various downstream tasks. However, relatively little is known about whether modern large language models (LLMs) can naturally attend to such expert-relevant features during sentiment classification tasks, particularly in zero-shot settings.

In this work, we investigate whether LLMs can benefit from expert-designed hints that highlight the importance of numbers in financial text. By comparing zero-shot prompts

with and without such hints, we find that explicitly reminding models to focus on numerical information significantly improves classification performance. This suggests that while LLMs are capable of processing complex text, they may overlook critical domain-specific cues unless explicitly guided. Our analysis also reveals that this performance gain is more pronounced in tweets that contain numerical data, indicating that expert-designed hints are especially helpful when relevant information is present but not automatically prioritized by the model. Interestingly, we also observe a phenomenon in cases where the sentiment expressed by the tweet’s writer differs from that perceived by external annotators. In these cases, adding expert-designed hints often leads to greater performance improvements, suggesting that such prompts may incidentally encourage more nuanced interpretations—potentially resembling a form of perspective-taking.

To better understand these dynamics, we conduct experiments using the Fin-SoMe dataset (Chen, Huang, and Chen 2020), which includes approximately 10k financial tweets annotated for both writer and reader sentiment. We evaluate four widely used LLMs—PaLM 2, Gemini Pro, GPT-3.5, and GPT-4—under different prompting strategies.

This study addresses the following research questions:

RQ1: Should LLMs be prompted to consider hints identified by experts, or are they inherently capable of recognizing such hints?

RQ2: Are the benefits of expert-designed hints more pronounced in tweets that contain numerical data?

RQ3: How do different categories of numbers (e.g., monetary vs. temporal) influence model performance when hints are used?

Related Work

Numbers have often been highlighted when dealing with financial documents. Some studies attempt to extract numerical information (Chen et al. 2019a,b). Others focus on the role of numbers in financial documents to explore reasoning skills (Zhu et al. 2021; Chen et al. 2021; Nan et al. 2022). Further research suggests that attention to numbers in financial documents can enhance downstream task performance, such as volatility forecasting (Yang et al. 2022; Shi et al.

Writer Sentiment	Whole	Perspective-Taking
Bullish	8,573	1,557
Bearish	1,427	277
Total	10,000	1,834

Table 1: Dataset Statistics.

2023). However, the relationship between sentiment analysis and numbers in text is rarely discussed. This paper explores whether LLMs still benefit from expert-proposed hints. Our findings suggest that a simple reminder can significantly improve performance in financial sentiment analysis, indicating that expert guidance may remain valuable in domain-specific tasks.

Experimental Setup

To evaluate whether LLMs can benefit from expert-designed hints in financial sentiment classification, we adopt three prompting strategies. The first is a simple zero-shot prompt, where the model is asked to classify sentiment without any reasoning or additional instructions. The second is the Chain-of-Thought (CoT) prompt (Wei et al. 2022), which guides the model to generate an intermediate reasoning process before outputting a final label. The third strategy, CoT + Hint, extends the CoT prompt by appending an expert-designed instruction: *“Focus particularly on any numerical data or statistics present in the tweet, as these figures may be crucial in determining the overall sentiment.”* This prompt aims to direct the model’s attention toward information that financial experts consider particularly salient. All prompts are applied uniformly across experiments, with no prompt tuning or in-context examples. The models are evaluated in purely zero-shot settings to assess their out-of-the-box reasoning and interpretive abilities.

The task is defined as a binary classification problem: given a tweet, the model must predict the writer’s intended sentiment—bullish or bearish. Although the dataset also contains reader-annotated sentiment labels, the classification target throughout our experiments is the writer-provided label. Tweets that received a “none” label from readers are retained, except where specific subset analyses apply. Model predictions are normalized to one of two valid labels, with light post-processing applied to ensure consistency across output formats.

We evaluate four state-of-the-art LLMs in this study. PaLM 2 and Gemini Pro were accessed in December 2023. GPT-3.5 and GPT-4 were accessed via the OpenAI API, specifically using versions “gpt-3.5-turbo-0125” and “gpt-4-0613”, respectively. All models were queried with deterministic settings (temperature = 0.0). The dataset statistics is presented in Table 1. This research does not involve training models; the entire dataset is treated as a test set to assess models’ proficiency in financial sentiment analysis.

As we treat the full dataset as a test set, performance is reported using two metrics: Micro-F1, which measures overall classification accuracy, and Weighted-F1, which accounts for class imbalance. To test for statistical significance be-

tween different prompting strategies, we apply McNemar’s test ($\alpha = 0.05$), with comparisons centered on whether CoT + Hint yields a significant improvement over the other two prompt types.

To better understand model behavior under challenging or information-sensitive conditions, we define two subsets of the data. The first is the perspective-taking subset, which includes tweets where the sentiment label assigned by the writer differs from that assigned by the reader (e.g., bullish vs. none). These cases often require deeper contextual understanding and may benefit from prompts that draw attention to latent cues. The second subset includes tweets containing at least one numerical expression. We identify these using regular expressions that capture integers, decimals, percentages, currency symbols, and other common numerical formats.

Finally, to address RQ3, we cross-reference the Fin-SoMe dataset with the FinNum dataset (Chen et al. 2019a), which provides fine-grained annotations for number types in financial text. From this alignment, we extract a subset of 6,493 tweets that are present in both datasets and use this set to analyze how different categories of numbers—such as monetary, temporal, and percentage—interact with prompt strategies in affecting model performance.

Experimental Results

Overall Performance (RQ1)

Table 2 presents the results across all prompting strategies and models. Among the four LLMs evaluated, PaLM 2 achieves the highest baseline performance under the zero-shot prompt, indicating strong out-of-the-box alignment with financial sentiment labels. However, its performance does not significantly improve with Chain-of-Thought (CoT) prompting, and the gain from adding hints is marginal. In contrast, for Gemini Pro, GPT-3.5, and GPT-4, we observe consistent and substantial improvements when the expert-designed hint is incorporated. Across these models, CoT prompting alone offers modest gains over zero-shot prompts, but the addition of the hint significantly enhances classification accuracy. Notably, in three out of four models, CoT + Hint achieves the best performance overall, with improvements statistically significant under McNemar’s test.

These findings imply that while LLMs are capable of performing financial sentiment analysis, they do not reliably attend to important numerical details unless explicitly prompted to do so. The observed gains suggest that expert-designed hints play a critical role in aligning model attention with domain-relevant information that might otherwise be underutilized.

Although not the primary focus of our study, we observe an intriguing pattern when analyzing tweets where the sentiment labels from writers and readers diverge. These “perspective-taking” cases are more challenging overall, with all prompting strategies yielding lower performance compared to the full dataset. However, the inclusion of the expert-designed hint once again leads to the largest performance gains in this subset. This suggests that even a simple numerical focus may act as a cue that helps the model inter-

LLM	Prompting Strategy	Full Dataset		Numeric Subset		Perspective-Taking Subset	
		Micro-F1	Weighted-F1	Micro-F1	Weighted-F1	Micro-F1	Weighted-F1
PaLM 2	Zero-shot	80.84	84.49	80.61	84.29	70.77	76.96
	CoT	79.09	83.06	80.02	83.57	68.97	75.63
	CoT + Hint	80.38	83.88	81.32*	84.39*	72.90*	78.30*
Gemini Pro	Zero-shot	68.26	77.48	69.06	78.01	42.80	55.71
	CoT	71.94	80.71	72.78	81.24	51.31	64.54
	CoT + Hint	74.45*	81.66*	75.32*	82.18*	54.20*	65.91*
GPT-3.5	Zero-shot	68.43	77.43	69.23	77.95	41.44	54.88
	CoT	68.99	77.75	69.80	78.27	43.35	55.76
	CoT + Hint	79.68*	83.72*	80.62*	84.23*	65.27*	73.22*
GPT-4	Zero-shot	78.01	84.59	78.93	85.13	54.80	67.68
	CoT	73.35	82.04	74.21	82.58	48.64	62.57
	CoT + Hint	81.86*	86.27*	82.82*	86.79*	62.00*	72.11*

Table 2: Performance across three prompting strategies on the full dataset, numeric subset, and perspective-taking subset. **Bold** values indicate the best result for each model. Asterisk (*) denotes statistically significant improvement over the second-best method using McNemar’s test ($p < 0.05$).

Category	Instance (%)	PaLM 2	Gemini Pro	GPT-3.5	GPT-4
Monetary	37.53	11.98	42.35	59.37	22.64
Temporal	30.23	0.80	18.16	32.85	7.74
Percent	13.32	-1.38	5.72	11.16	7.79
Quantity	12.46	1.48	9.77	14.83	4.20
Indicator	2.43	0.00	0.00	10.76	4.43
Option	2.28	-8.98	8.98	25.17	2.99
Product Number	1.74	1.77	10.62	17.70	13.27

Table 3: Improvement (%) in the subset of tweets containing a number in the target category.

pret underlying intentions more effectively. While we do not claim to model true perspective-taking in the psychological sense, this effect warrants further investigation.

Tweets Containing Numbers (RQ2)

To further test whether the expert hint’s benefit is tied to numerical content, we evaluate model performance on a subset of tweets that include at least one number. The results in Table 2 reveal a clear pattern: all LLMs achieve higher scores in this numeric subset when the CoT + Hint prompt is used. In fact, performance with CoT + Hint on numeric tweets consistently exceeds performance on the full dataset across all models.

This supports RQ2 by confirming that expert-designed hints are especially effective in contexts where numbers are present. Interestingly, for models like PaLM 2 that showed only marginal benefit from hinting on the full dataset, the gains become more pronounced within this numeric subset. This suggests that LLMs may not automatically identify numbers as semantically important in financial sentiment classification—even when those numbers are present—unless explicitly instructed to do so.

Effect of Number Category (RQ3)

To answer RQ3, we examine how the type of number present in a tweet influences the effectiveness of the expert-designed hint. Using aligned Fin-SoMe and FinNum data, we group

tweets by number category and measure improvement in Micro-F1 from the zero-shot prompt to CoT + Hint.

As shown in Table 3, tweets containing monetary values constitute the largest group (37.5%) and also exhibit the most substantial performance gains, especially for GPT-3.5 and Gemini Pro. Improvements are also observed in other categories such as temporal and quantity, though to a lesser extent. While a few categories (e.g., option numbers in PaLM 2) show negative deltas, the overall trend is consistent: hints help most when the number is financially meaningful. These results indicate that the effectiveness of expert hints depends not only on the presence of numbers, but also on the semantic role those numbers play in the financial context. Prompts that draw attention to these figures help models align with human-like heuristics used in interpreting financial sentiment.

Limitations and Future Directions

Firstly, the findings of this study are primarily based on financial social media data, particularly from the Stocktwits platform. This focus may limit the generalizability of our conclusions to other domains or types of social media content. Future studies could explore whether the observed benefits of perspective-taking and expert-designed hint extend to other domains, such as healthcare or politics, where sentiment analysis is equally critical.

Secondly, this study simplifies the concept of perspective-

taking. However, perspective-taking in human communication is a complex, multi-dimensional process that involves understanding emotional states, intentions, and contextual factors. Future work could aim to model these additional layers of complexity to achieve a more holistic understanding of sentiment in social media texts.

Another limitation is the focus on only four LLMs in our experiments. While these models are among the most advanced at the time of our study, the rapidly evolving field of natural language processing continually introduces new models that may offer different insights into the challenges of financial sentiment analysis. Testing our approach with a wider array of LLMs could provide a more comprehensive understanding of its effectiveness.

Lastly, our study's focus on numerical data as a key element of financial sentiment analysis may overlook other important factors that influence sentiment interpretation, such as linguistic subtleties, cultural references, or domain-specific knowledge. Incorporating these dimensions into future research could provide a more holistic understanding of the challenges and opportunities in applying large language models to financial sentiment analysis.

Conclusion

We investigated whether expert-designed hints can improve financial sentiment analysis with large language models. Our results show that explicitly prompting models to attend to numerical information leads to consistent performance improvements, particularly on tweets containing numbers or requiring deeper interpretation. These findings suggest that even strong LLMs may overlook domain-relevant cues unless guided, and that simple, targeted prompts can enhance their effectiveness in specialized tasks.

References

- Baker, M.; and Wurgler, J. 2007. Investor sentiment in the stock market. *Journal of economic perspectives*, 21(2): 129–151.
- Chen, C.-C.; Huang, H.-H.; and Chen, H.-H. 2018. NTUSD-Fin: a market sentiment dictionary for financial social media data applications. In *Proceedings of the 1st financial narrative processing workshop (FNP 2018)*, 37–43.
- Chen, C.-C.; Huang, H.-H.; and Chen, H.-H. 2020. Issues and Perspectives from 10,000 Annotated Financial Social Media Data. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6106–6110. ISBN 979-10-95546-34-4.
- Chen, C.-C.; Huang, H.-H.; Shiue, Y.-T.; and Chen, H.-H. 2018. Numeral understanding in financial tweets for fine-grained crowd-based forecasting. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 136–143. IEEE.
- Chen, C.-C.; Huang, H.-H.; Takamura, H.; and Chen, H.-H. 2019a. Overview of the ntcir-14 finnum task: Fine-grained numeral understanding in financial social media data. In *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*, 19–27.
- Chen, C.-C.; Huang, H.-H.; Tsai, C.-W.; and Chen, H.-H. 2019b. Crowdpt: Summarizing crowd opinions as professional analyst. In *The World Wide Web Conference*, 3498–3502.
- Chen, Z.; Chen, W.; Smiley, C.; Shah, S.; Borova, I.; Langdon, D.; Moussa, R.; Beane, M.; Huang, T.-H.; Routledge, B.; and Wang, W. Y. 2021. FinQA: A Dataset of Numerical Reasoning over Financial Data. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3697–3711. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.
- Liu, S. 2015. Investor sentiment and stock market liquidity. *Journal of Behavioral Finance*, 16(1): 51–67.
- Nan, L.; Hsieh, C.; Mao, Z.; Lin, X. V.; Verma, N.; Zhang, R.; Kryściński, W.; Schoelkopf, H.; Kong, R.; Tang, X.; Mutuma, M.; Rosand, B.; Trindade, I.; Bandaru, R.; Cunningham, J.; Xiong, C.; Radev, D.; and Radev, D. 2022. FeTaQA: Free-form Table Question Answering. *Transactions of the Association for Computational Linguistics*, 10: 35–49.
- Shi, M.-X.; Chen, C.-C.; Huang, H.-H.; and Chen, H.-H. 2023. Enhancing Volatility Forecasting in Financial Markets: A General Numeral Attachment Dataset for Understanding Earnings Calls. In Park, J. C.; Arase, Y.; Hu, B.; Lu, W.; Wijaya, D.; Purwarianti, A.; and Krisnadhi, A. A., eds., *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, 37–42. Nusa Dua, Bali: Association for Computational Linguistics.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.
- Xu, Y.; and Cohen, S. B. 2018. Stock Movement Prediction from Tweets and Historical Prices. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1970–1979.
- Yang, L.; Li, J.; Dong, R.; Zhang, Y.; and Smyth, B. 2022. NumHTML: Numeric-Oriented Hierarchical Transformer Model for Multi-task Financial Forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11604–11612.
- Zhu, F.; Lei, W.; Huang, Y.; Wang, C.; Zhang, S.; Lv, J.; Feng, F.; and Chua, T.-S. 2021. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. 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)*, 3277–3287. Online: Association for Computational Linguistics.