International Conference on Sustainable Economics, Management, and Accounting (ICSEMA 2025)
Proceedings
Vol. 01 No.01 Year 2025, Page 1755-1763

ANALYSIS OF PROMPT ENGINEERING EFFECTIVENESS IN STOCK RECOMMENDATION BY CHATGPT: AN EXPERIMENTAL STUDY IN THE INDONESIAN MARKET

Hadi Setyo Nugroho[1*], Intan Shaferi[2]

[1,2] Management/Faculty of Economic and Business, Universitas Jenderal Soedirman, Indonesia
[*]Email corresponding author: hadinrh99@gmail.com

Abstract

This study investigates the influence of prompt engineering on the quality of stock recommendations generated by ChatGPT in the Indonesian energy sector. Using four distinct prompt types ranging from general to highly structured the research analyzes outputs related to ten IDX-listed energy stocks. Each ChatGPT response was evaluated using four binary-coded indicators: analytical depth, indicator integration, scenario contextualization, and actionability. The findings reveal that structured and specific prompts produce significantly more accurate, relevant, and actionable recommendations. Among all prompt types, time-bound and context-rich prompts delivered the highest performance, while vague prompts yielded generic, low-quality outputs. The results support the importance of prompt literacy and suggest that effective human-AI interaction in financial decision-making depends heavily on input clarity. This study contributes to the growing literature on generative AI in finance and highlights the need for user education in prompt design.

Keywords: ChatGPT, Prompt Engineering, Stock Recommendation, Content Analysis, Indonesia Stock Market

INTRODUCTION

Artificial Intelligence (AI) is increasingly transforming the landscape of financial decision-making, particularly in the domain of investment analysis. Among recent innovations, Large Language Models (LLMs) like ChatGPT, developed by OpenAI, have gained significant traction due to their accessibility and conversational capabilities. These tools allow users especially retail investors to obtain rapid insights on stock performance, technical analysis, and market sentiment (Chen *et al*., 2022; Brown et al., 2020). Despite their promise, however, LLMs are inherently non-deterministic, meaning that the quality and direction of their outputs heavily depend on the formulation of input queries, commonly referred to as prompt engineering.

Prompt engineering has emerged as a crucial mediator between human input and AI output quality. Unlike traditional analytical software, which relies on fixed inputs and algorithms, LLMs respond based on probabilistic token generation, making them highly sensitive to the phrasing, structure, and context embedded in user prompts (Reynolds & McDonell, 2021; Zhang *et al*., 2023). In financial applications, this variability introduces risk, particularly when vague or unstructured prompts lead to generic or even misleading recommendations (Zhao *et al*., 2023). In high-stakes environments such as equity trading, prompt clarity and specificity become vital in guiding AI responses toward analytical depth and actionable insight.

On the theoretical front, prompt engineering raises important debates: Can prompt design achieve deterministic improvements across LLM models, or is it fundamentally brittle and reliant on heuristic experimentation? Critics argue that the field still lacks formal rigor and standardization. Additionally, in financial reasoning tasks, it remains an open question whether structured prompts alone (contextual framing, temporal focus, persona prompting) are sufficient to overcome

limitations like hallucination, or if deeper techniques such as retrieval-augmented generation are required for robust output quality. While prompt engineering is gaining traction, there remains debate on whether strategies like Chain of Thought (CoT) or Retrieval-Augmented Generation (RAG) are necessary for more reliable financial reasoning. CoT enables step-by-step logic formulation, improving transparency, while RAG helps reduce hallucination by grounding outputs in retrieved factual data (Wei et al., 2022; Lewis et al., 2020).

Recent studies support the notion that prompt design significantly impacts the output relevance in financial text generation. Hwang *et al*. (2023) argue that prompt literacy the ability to formulate, interpret, and iterate prompts is now a necessary digital skill in the age of AI-assisted financial tools. However, there remains a gap in empirical understanding of how prompt variation tangibly affects the quality of AI-generated investment recommendations. While prior studies have examined LLMs in static forecasting scenarios, few have experimentally tested the variance of model outputs under prompt manipulation in real-time financial contexts, particularly in emerging markets such as Indonesia. This study aims to fill that gap by analyzing how ChatGPT's stock recommendations change when different prompt types are applied across selected energy-sector equities on the Indonesia Stock Exchange (IDX).

By integrating content analysis with a multi-prompt experimental framework, this research contributes to the growing discourse on AI-human collaboration in financial decision-making. Specifically, we aim to (1) quantify the influence of prompt specificity and structure on the quality of recommendations, and (2) promote investor education on the importance of prompt engineering in interacting with generative AI models.

LITERATURE REVIEW AND HYPOTHESIS FORMULATION

Prompt Engineering

Prompt engineering refers to the deliberate construction of input queries to optimize responses generated by large language models (LLMs), such as ChatGPT (Reynolds & McDonell, 2021). In financial contexts, the structure and specificity of prompts play a critical role in determining the analytical depth, contextual relevance, and clarity of the output. This concept is theoretically grounded in the Input–Output Theory (Shannon & Weaver, 1949), which posits that the precision of input directly influences the quality of output in communication systems, including AI-based models.

Zhao *et al*. (2023) empirically demonstrated that specific and well-structured prompts significantly enhance the relevance and clarity of financial texts generated by generative AI. Similarly, Chen *et al*. (2022) observed that while LLMs have shown promising capabilities in financial forecasting, their performance often deteriorates when prompts lack contextual richness or are overly generalized. This underlines the importance of precise prompt formulation in guiding the AI's generative process.

Brown *et al*. (2020) further explain that LLMs operate based on probabilistic token generation rather than deterministic logic. As such, they are not inherently analytical but rely heavily on the direction provided through structured inputs. In this sense, the quality of human-AI interaction is largely mediated by the clarity, depth, and focus of the prompt.

Prompt Literacy

In investment analysis, effective prompts may incorporate elements such as technical indicators, investment time horizons, macroeconomic trends, and investor profiles. Hwang *et al*. (2023) introduce the concept of prompt literacy, a new form of digital literacy defined as the ability to formulate, refine, and interpret prompts effectively in order to receive meaningful and actionable AI outputs. Their study on AI adoption in finance revealed that users who iteratively refine their prompts tend to receive more robust and practically useful investment recommendations.

In the context of retail investing, where users increasingly depend on AI-based tools for financial decision-making, understanding how to engineer prompts becomes a vital skill. Poorly constructed prompts may yield vague or misleading outputs, while well-crafted ones can result in

responses that are informative, targeted, and suitable for decision-making. Therefore, prompt engineering is not merely a technical procedure but a strategic element of AI interaction that has significant implications for user outcomes in the financial sector.

Hypothesis Formulation

Decision-making theories in AI-assisted environments emphasize that the quality of system output is highly influenced by the nature of user input. In the context of ChatGPT-generated stock recommendations, prompt formulation serves not just as a command but as a guide that shapes the model's analytical trajectory. A well-structured prompt provides clear direction, enabling the model to deliver responses that are contextually accurate and actionable.

Zhao *et al*. (2023) highlight that structured prompts improve output consistency and contextual integration, particularly in financial applications. Hwang *et al*. (2023) also argue that prompt literacy, the ability to generate precise, goal-oriented inputs, is emerging as a key digital competency. Their findings suggest that prompt engineering is directly linked to the effectiveness of LLM-generated financial advice, especially in markets characterized by volatility and information asymmetry, such as Indonesia's energy sector.

Moreover, content analysis approaches such as those proposed by Krippendorff (2018) allow researchers to evaluate AI-generated content using standardized indicators like analytical depth, integration of financial data, contextual relevance, and clarity of recommendation. In this study, prompt variations were systematically applied to energy-sector stocks on the Indonesia Stock Exchange (IDX), and responses were analyzed using binary-coded evaluation criteria. Results showed that prompts with clearly defined timeframes, relevant financial indicators, and market context scored significantly higher in all quality dimensions.

These empirical findings reinforce the theoretical assumption that the quality of AI-generated stock recommendations is not solely determined by the model's internal capabilities but also by the sophistication of user interaction, specifically, prompt design. For retail investors relying on AI to inform financial decisions, acquiring prompt literacy is not optional but essential.

This literature leads to form several hypotheses to be tested in the study:

H0 : There is no significant difference in the accuracy and consistency of stock recommendations generated by ChatGPT across different prompt structures.

H1 : Structured and specific prompts will generate more accurate and consistent stock recommendations compared to general and vague prompts.

RESEARCH METHODS

This research adopts a quantitative experimental design aimed at exploring how variations in prompt structure influence the quality and consistency of ChatGPT-generated stock recommendations. The study utilizes a single-factor, multi-level treatment approach, wherein prompt types act as the independent variable, and quality indicators serve as dependent variables. Output evaluation was carried out using a structured content analysis method with four binary-coded metrics.

The population for this research comprises all listed stocks on the Indonesia Stock Exchange (IDX). To maintain relevance with the financial context and ensure high output consistency from ChatGPT, the sample was focused on the energy sector, a sector with strategic importance in Indonesia's economic landscape and a frequent subject of investor analysis.

An expanded purposive sampling technique was applied to select ten energy-sector companies based on the following criteria:

1. High liquidity: measured through average daily trading volume over the past 6 months.
2. Market relevance: companies included in major indices such as the IDXENERGY or LQ45.
3. Diversity of sub-sectors: to include a mix of oil & gas, renewables, power generation, and integrated energy services.

4. Data availability: sufficient financial data and news coverage to allow ChatGPT to generate meaningful outputs.

Data were collected by interacting with ChatGPT (GPT-4 model), developed by OpenAI. For each of the ten stocks, four different prompt formats were submitted. The responses were recorded and used SPSS for statistical analysis. Descriptive statistics were calculated to determine mean performance per prompt type. A one-way ANOVA and non-parametric test were employed to assess whether differences across prompt types were statistically significant. This mixed-method approach enhances objectivity and replicability in evaluating AI-generated financial content.

1. Prompt A: General inquiry
   Based on the current market conditions and overall performance of PT Perusahaan Gas Negara Tbk (PGAS), is this stock considered a good long-term investment for retail investors in Indonesia? Please include a general assessment of the company's business prospects, competitive position, and any relevant market trends influencing your recommendation.
2. Prompt B: Technical indicator-based
   Please provide a technical analysis of PGAS using key indicators. Interpret the current trend direction, potential support and resistance levels, and indicate whether the current momentum suggests a buy, sell, or hold signal for short-term traders.
3. Prompt C: Contextual and timeframe specific
   Considering PGAS's historical price action, trading volume, and external macroeconomic factors (e.g., global gas prices, Indonesian energy policies), what is your 2-week outlook for this stock? Please explain the short-term trend projection, including expected volatility, possible catalysts, and whether the current level is favorable for entry or exit positions.
4. Prompt D: Actionability and clarity analysis
   As a new retail investor considering PGAS for portfolio inclusion, can you evaluate the key investment risks and potential opportunities? Please address market risk, regulatory/policy risks, company-specific risks (e.g., debt, revenue concentration), and growth opportunities in Indonesia's gas sector. Conclude with an investment suitability rating for conservative investors.

Each of the 40 resulting outputs (10 stocks × 4 prompts) was recorded and analyzed using content analysis based on the following four binary-coded indicators (Krippendorff, 2018):

1. Analytical Depth and Justification
   Measures the extent to which the ChatGPT response includes explicit and well-structured reasoning behind its investment recommendation, whether based on quantitative data (e.g., financial ratios, trends) or qualitative insights (e.g., business strategy, sector outlook).
   Score 0: The answer is vague or generic, lacking justification or structured argumentation.
   Score 1: The response includes clear reasoning supported by data or logical interpretation relevant to the recommendation.
2. Indicator Integration and Accuracy
   Assesses whether financial or technical indicators (e.g., RSI, MACD, moving averages, debt ratio) are not only mentioned but also accurately interpreted and contextually relevant to the stock being analyzed.
   Score 0: Indicators are incorrectly interpreted or merely listed without meaningful integration.
   Score 1: Indicators are applied appropriately, interpreted correctly, and integrated into the recommendation.
3. Scenario Contextualization
   Evaluates the response's ability to incorporate relevant external or internal factors, such as macroeconomic conditions, policy developments, or investor profiles (e.g., risk appetite).
   Score 0: No meaningful context is provided or context is generic and not tied to PGAS.
   Score 1: The response effectively integrates contextual elements and links them to the analysis.

4. Actionability and Clarity of Output

   Measures whether the response offers clear, logical, and actionable insights that investors can use for decision-making. This includes structured reasoning, clarity of recommendation, and usability (e.g., entry point, timeframe, risk exposure).

   Score 0: Response is unclear, lacks structure, or fails to offer a practical conclusion.

   Score 1: Response is well-organized, clear, and contains recommendations that can guide investor action.

Each output received a total score ranging from 0 to 4. These binary scores were then averaged across five stocks for each prompt type, resulting in a proportion of fulfilled criteria per category. For example, a score of 0.60 under "Clear Recommendation" indicates that 3 out of 5 outputs using that prompt explicitly stated buy/sell/hold recommendations. This method allows for cross-comparison between prompt structures in an objective, measurable format.

To improve validity, scoring criteria were adapted from prior content analysis literature (Krippendorff, 2018). Binary indicators were strictly defined to reduce subjectivity. The researcher ensured inter-item consistency by rechecking scores independently over two sessions. While no external expert was used, consistency checks and documentation of responses aimed to uphold internal reliability. O'Connor & Joffe (2020) emphasize that inter-coder reliability (ICR) is essential for ensuring consistency in qualitative content coding. While this study did not employ multiple coders, repeated evaluations by the same researcher were used to increase internal consistency.

Furthermore, Bolognesi *et al*., (2017) in *Behavior Research Methods* underscore that reliability in content analysis must account for coding stability, replicability, and accuracy. This study followed their suggestion by using binary-coded, explicitly defined indicators to reduce subjective interpretation. Additionally, Dwivedi *et al*., (2021) highlights that Krippendorff's alpha is the most appropriate measure of agreement when coding natural language content particularly when using dichotomous scales, as applied in this study. While Krippendorff's alpha was not calculated due to the single-coder nature of this research, the scoring framework is designed to be reproducible in future studies employing multiple raters. Overall, these procedures were applied to strengthen the methodological rigor and reproducibility of the analysis, while acknowledging the study's current limitations in inter-coder reliability assessment.

RESULTS AND DISCUSSION

The results of this study clearly demonstrate that the structure and specificity of prompts significantly influence the quality of stock recommendations generated by ChatGPT. The experiment applied four types of prompts (A–D) to five energy-sector stocks listed on the IDX, with each output evaluated using a binary content analysis rubric. The evaluation was based on four indicators: clarity of recommendation, use of financial indicators, contextual relevance, and alignment with actual stock movement over a two-week window.

Table 1. Content Analysis Scoring by Prompt Type (Proportion of Fulfilled Criteria)

Prompt Type	Analytical Depth and Justification	Indicator Integration and Accuracy	Scenario Contextualization	Actionability and Clarity	Average Score (0–4)
Prompt A	0.20	0.30	0.20	0.30	1.00
Prompt B	0.70	0.70	0.40	0.70	2.50
Prompt C	0.80	0.90	0.90	0.70	3.30
Prompt D	0.60	0.90	0.70	0.80	3.00

Table 2. Descriptive Statistics of Total Score per Prompt Type

Prompt Type	Mean	Std. Deviation	Minimum	Maximum
A (General Inquiry)	1.00	1.155	0	3
B (Technical Analysis)	2.50	0.707	1	3
C (Contextual + Timeframe)	3.30	0.483	3	4
D (Action-Oriented)	3.00	0.816	2	4

The results strongly reject the null hypothesis (H0) and support the alternative hypothesis (H1): that structured and specific prompts lead to more accurate and consistent outputs compared to general and vague ones. This is evident from the scoring results shown in Table 1, where Prompt C which contained timeframe and contextual elements consistently outperformed other prompt types, achieving an average total score of 3.30 out of 4. In contrast, Prompt A, which was general and nonspecific, scored the lowest average (1.00), indicating poor recommendation clarity and lack of actionable content.

Table 3. ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	31.300	3	10.433	15.268	.000
Within Groups	24.600	36	.683		
Total	55.900	39			

Table 4. Kruskall-Wallis Test

	Prompt	N	Mean Rank
Total Score	A	10	8.80
	B	10	19.45
	C	10	28.75
	D	10	25.00
	Total	40	

Table 5. Test Statistic

	Total Score
Kruskal-Wallis H	18.739
df	3
Asymp. Sig.	.000

A One-Way ANOVA was conducted to assess whether observed differences in total scores between prompt types were statistically significant. The results were conclusive:

1. $F(3,36) = 15.27$, $p < 0.001$, confirming significant variation between prompt structures.
2. A Kruskal-Wallis test, as a non-parametric validation, also returned a significant result ($H = 18.74$, $p < 0.001$) indicating robust consistency across statistical models.

The descriptive statistics indicate a clear performance disparity across prompt types. Prompt C, which incorporated contextual and time-bound information, yielded the highest average total score (Mean = 3.30, SD = 0.48), demonstrating consistent and high-quality responses. This was followed by Prompt D (M = 3.00), which emphasized risk-awareness and investor positioning. Meanwhile, Prompt A, characterized by vague and general inquiries, recorded the lowest average score (M = 1.00) with high variance (SD = 1.15), indicating poor and inconsistent recommendation quality. These findings affirm that prompt specificity significantly enhances the analytical capability of ChatGPT, aligning with prior research emphasizing prompt literacy as a key factor in generative AI usage in finance (Hwang et al., 2023).

These findings show that Prompt C performs best across all indicators, particularly in contextual relevance and recommendation clarity. This aligns with the findings of Zhao *et al*. (2023), who emphasized that well-structured prompts significantly increase output relevance in LLM-generated financial texts. Moreover, the use of content analysis removes dependency on human raters' subjective perceptions, making the results replicable and measurable, as recommended by Krippendorff (2018).

Table 6. Example ChatGPT Responses and Scoring

Prompt Type	Stock	Excerpt of ChatGPT Output	Score [Anjust[1], Ind[2], Scene[3], Act[4]]
A	PGAS	"PGAS is a significant player in Indonesia's energy sector. It could be a good investment in the long run."	[0, 0, 0, 0]
B	ELSA	"MACD indicates bullish momentum; RSI is currently at 47. Immediate resistance is around 420."	[1, 1, 0, 1]
C	SURE	"Given volume spikes, global oil recovery, and 2-week moving average, SURE likely to trend upward near term."	[1, 1, 1, 1]
D	MEDC	"MEDC's moderate debt level and regulatory compliance make it viable for conservative investors seeking stability."	[1, 0, 1, 1]

Notes: Analytical Depth and justification[1], Indicator Integration[2], Scenario Contextualization[3], Actionability and Clarity[4]

1. **Prompt C** outperformed others in all dimensions, with an average total score of 3.5 out of 4.
2. **Prompt A** frequently failed to provide actionable insight or timeframe-based analysis.
3. **Prompt B** consistently integrated financial indicators but lacked contextual framing.
4. **Prompt D** showed strength in investor-type relevance but had lower price alignment

The consistent outperformance of Prompt C demonstrates the importance of incorporating timeframes, historical patterns, and macro variables into prompt design. This finding echoes Brown et al. (2020), who note that LLMs like ChatGPT respond more meaningfully to structured, well-framed inputs due to their probabilistic nature. Similarly, Prompt B performed well in integrating technical indicators, but often lacked scenario context. Prompt D emphasized risk and investor suitability, performing strongly on contextualization and depth, though slightly weaker in indicator integration.

These findings align with Krippendorff (2018), who emphasizes the role of clearly defined evaluation categories in achieving reliable and interpretable text content analysis. The binary evaluation used here allowed objective scoring across prompts and reduced interpretive bias. Furthermore, the results reinforce the concept of prompt literacy (Hwang *et al*., 2023), defined as the user's ability to formulate effective queries to maximize AI response relevance. Mistry (2025) also found that generative AI only benefits financial education when prompts are structured and specific.

The findings confirm the arguments made by Zhang *et al*., (2023), who noted that prompt specificity enhances output relevance. Furthermore, our results align with Brown et al., (2020), who stated that LLMs are not inherently analytical but reactive to structured guidance. Even though ChatGPT has access to a broad corpus of financial knowledge, it requires clear guidance to generate valuable outputs. The experiment also revealed that overly vague prompts often led to generic responses lacking actionable insights, consistent with observations by Reynolds & McDonell (2021).

These results support the hypothesis and suggest that investor reliance on LLMs must be coupled with prompt literacy. In line with that, Hwang *et al*. (2023) define prompt literacy as "the ability to generate precise prompts interpret the outputs, and iteratively refine prompts to achieve desired results". Consequently, inaccurate or ambiguous queries may mislead rather than aid investment decisions, highlighting the need to integrate prompt engineering education into financial literacy programs to enhance investor outcomes.

CONCLUSION

This study concludes that prompt engineering significantly influences the quality of stock recommendations generated by ChatGPT. Structured and specific prompts particularly those incorporating clear timeframes, financial indicators, and investor context produced more coherent, consistent, and actionable insights. Among the four prompt types tested, Prompt C yielded the highest overall performance across all evaluation indicators.

The findings imply that the effectiveness of AI-based financial tools is not solely determined by model capability but also by user interaction quality. Practically, this means investors, especially retail participants, must be equipped with prompt literacy skills to fully leverage AI-powered platforms like ChatGPT. Financial educators and training programs should consider embedding prompt engineering modules within digital financial literacy curricula. Platforms that provide AI-based advisory features can also benefit from implementing guided prompt templates to minimize vague or misleading queries.

This study has several limitations. First, it only used one LLM model (ChatGPT), which may limit generalizability to other generative AI systems such as Claude, Gemini, or LLaMA. Second, the analysis focused exclusively on companies in the energy sector, meaning the results might not be representative across different industries. Third, the evaluation of output relied on a content analysis of five binary indicators, which, while objective, may still overlook nuance in natural language interpretation.

Future studies are encouraged to adopt a comparative approach by including multiple LLMs or integrating ensemble models to determine prompt effectiveness across systems. Expanding the scope beyond energy sector stocks to other industries or markets can offer broader insight into the universality of the prompt engineering effect. Additionally, longer-term tracking of prediction accuracy over monthly or quarterly horizons can provide a more robust measure of alignment between AI recommendations and actual stock movements. Finally, future work could explore the use of semi-automated scoring systems using natural language processing (NLP) tools to improve scalability and reproducibility in prompt evaluation.

BIBLIOGRAPHY

Bolognesi, M., Pilgram, R., & Van Den Heerik, R. (2017). Reliability in content analysis: The case of semantic feature norms classification. *Behavior Research Methods*, *49*, 1984-2001.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Dwivedi, Y. K., Ismagilova, E., Hughes, D. L., Carlson, J., Filieri, R., Jacobson, J., ... & Wang, Y. (2021). Setting the future of digital and social media marketing research: Perspectives and research propositions. *International journal of information management*, *59*, 102168.

Hwang, Y., Lee, J. H., & Shin, D. (2023). What is prompt literacy? An exploratory study of language learners' development of new literacy skill using generative AI. *arXiv preprint arXiv:2311.05373*. https://doi.org/10.48550/arXiv.2311.05373

Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology* (4th ed.). Sage Publications.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, *33*, 9459-9474.

Mistry, H. (2025). Utilizing Generative AI for Financial Literacy. Journal of Computer Science and Technology Studies, 7(3), 253-261. 10.32996/jcsts.2025.7.3.28

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: debates and practical guidelines. *International journal of qualitative methods*, *19*, 1609406919899220.

Reynolds, L., & McDonell, K. (2021). Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. *arXiv:2102.07350*. https://doi.org/10.48550/arXiv.2102.07350

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837.

Zhang M., Jin L., Song L., Mi H., Chen W., and Dong Yu. 2023. SafeConv: Explaining and Correcting Conversational Unsafe Behavior. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22–35, Toronto, Canada. Association for Computational Linguistics. 10.18653/v1/2023.acl-long.2

Zhao, Z., & Welsch, R. E. (2024). Aligning LLMs with Human Instructions and Stock Market Feedback in Financial Sentiment Analysis. *arXiv preprint arXiv:2410.14926*. https://doi.org/10.48550/arXiv.2410.14926