Beyond Turing Test: Can GPT-4 Sway Experts' Decisions?

Anonymous Submission

Abstract

In the post-Turing era, evaluating large language models (LLMs) involves assessing generated text based on readers' reactions rather than merely its indistinguishability from human-produced content. This paper explores how LLM-generated text impacts readers' decisions, focusing on both amateur and expert audiences. Our findings indicate that GPT-4 can generate persuasive analyses affecting the decisions of both amateurs and professionals. Furthermore, we evaluate the generated text from the aspects of grammar, convincingness, logical coherence, and usefulness. The results highlight a high correlation between real-world evaluation through audience reactions and the current multi-dimensional evaluators commonly used for generative models. Overall, this paper shows the potential and risk of using generated text to sway human decisions and also points out a new direction for evaluating generated text, i.e., leveraging the reactions and decisions of readers. We release our dataset to assist future research.

1 Introduction

Large language models (LLMs) have demonstrated impressive performance, and the Turing test has become less reliable for evaluating LLM-generated text (Tikhonov and Yamshchikov, 2023). In other words, pursuing the generation of content indistinguishable from that produced by humans is no longer the goal in the post-Turing era. Nowadays, we should evaluate LLM-generated text using the same criteria applied to human-generated text. In the real world, these criteria are always related to readers' reactions. For example, the number of views is an important evaluation metric for YouTube videos, the number of likes is the evaluation metric for social media editors, and the obtained donations are the best metrics for crowdfunding proposals. Following this line of thought, this paper provides a pilot exploration of linking generated text with readers' reactions. Going a step further, the behaviors and reactions of common people and experts are very different (Snow et al., 2008; Aguda et al., 2024). To analyze this difference, we include the reactions of both amateurs and experts for in-depth discussions.

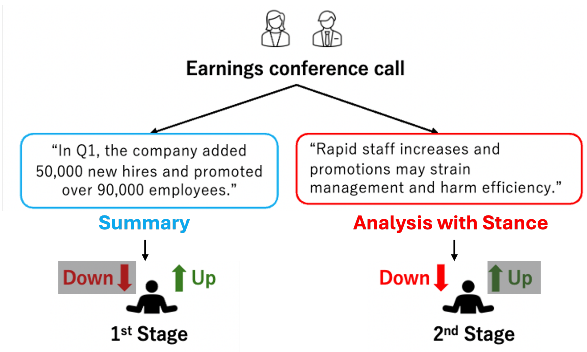

Figure 1: Design of experiments.

Inspired by previous studies (Kimbrough, 2005; Keith and Stent, 2019), earnings conference calls (ECCs)—meetings among company managers and professional analysts to discuss the latest operations and future plans—affect both amateur and professional investors' decisions. This scenario fits our scope, which aims to discuss how the information provided influences amateurs' and experts' decisions. Therefore, we designed our experiments based on ECCs. Figure 1 illustrates the design of the experiment. We first provide an objective summary of the ECC and ask investors to predict whether to increase or decrease based on the given summary. Then, we provide a subjective analysis for the same ECC to investors and ask them to decide whether they want to change their decisions. Our results reveal that GPT-4 (OpenAI, 2023) can generate persuasive analysis that sways both amateurs' and professionals' decisions.

Given that many recent studies (Zhong et al., 2022; Chan et al., 2023) propose evaluating generated text by scoring, we also assess the generated text from both objective (grammar) and subjective (convincingness, logical coherence, and usefulness) aspects. Our results indicate that both objective and subjective evaluation metrics do highly correlate with the decisions. The high correlation between multi-dimensional evaluators and real-world evaluations (audience/reader reactions) in our experiment highlights the potential of using readers' reactions as an evaluation method.

To sum up, this paper focuses on the following research questions:

(RQ1): To what extent does state-of-the-art LLM-generated text sway people's decisions?

(RQ2): Are the generated text's influences on amateurs and professionals different?

(RQ3): Does the recent popular evaluation approach align with reactions?

2 Related Work

The impact of text information on financial markets is a widely studied topic. Research has shown that different kinds of text data, from social media to financial news, can affect both trading algorithms and investor behavior (Karppi and Crawford, 2016; Arcuri et al., 2023). Furthermore, the effects of bullish articles, created as part of stock promotion schemes, have been examined for their ability to draw investor attention and influence the market (Clarke et al., 2020). The relationship between artificial intelligence and investor decision-making is another key area of research. Lai et al. (2023) reviews recent studies exploring how AI and humans interact in various domains, including finance. Additionally, research examines how machine learning results affect investor choices (Biran and McKeown, 2017). Despite considerable research into NLP applications in finance, the influence of text on financial markets, and the interaction between AI and investors, there remains a gap in studies specifically examining the impact of LLMs on investors' decisions. Our paper addresses this gap by proposing a novel evaluation framework.

3 Experimental Design

3.1 Dataset

We adopt the ECTSum (Mukherjee et al., 2022) dataset as the base for our experiment. In ECTSum, there are 2,425 ECC transcripts with professional journalist-written summaries. We manually aligned these data with the professional analysis reports on the Bloomberg Terminal,[1] which is one of the largest financial information vendor platforms. Finally, we obtained 234 instances containing the corresponding analysis reports. GPT-4 (OpenAI, 2023) [2] was used to generate the analysis by providing the ECC transcript and the stance (Overweight/Underweight), where overweight (underweight) denotes the suggesting increasing (decreasing) stock prices. Inspired by Kogan et al. (2023), providing analysis from a certain aspect is rational, but intent to promote the analysis from a certain aspect is illegal. Thus, in addition to having GPT-4 act as a professional analyst, we also had GPT-4 act as a promoter to render and write an analysis with a stronger stance.[3]

3.2 Evaluation Paradigm

We recruited five financial experts with over five years of industry experience and eight students with academic backgrounds in finance for the experiment. There are two stages in each round of the experiment. In the first stage, participants are presented with neutral summaries, either professional journalist-written or GPT-4-generated summaries. Participants are asked to decide whether to increase or decrease the stock of the company within three-day trading period following the conference date. In the second stage, participants received a document with an investment stance pertaining to the same ECC as in the first stage. The documents are either professional analysis reports or GPT-4-generated analyses with stance. They were again asked to make a decision for the same three-day period. Here, a three-day setting was selected based on the empirical study of previous work (Birru et al., 2022), which supports that the market reflects information within three days. In this way, we can answer (RQ1) and (RQ2) by analyzing the change between the two stages and the difference between students (amateurs) and experts (professionals).

The basic salary of participants is 180% of the minimum salary stipulated by law. To mimic real-world incentives and motivate participants to try their best to make the decision, their salary will increase to 270% of the minimum salary stipulated by law as a reward if they make the correct deci-

[1] https://www.bloomberg.com/professional/products/bloomberg-terminal/

[2] We utilize `gpt-4-1106` in our experiments.

[3] All prompts are available in Appendix A.

2nd Stage Source	All	Amateur	Expert	Veteran
GPT-4	28.7%	31.3%	24.7%	15.6%
Analyst	26.3%	25.0%	28.3%	21.2%

Table 1: Ratio of changing decisions in the second stage.

Change	Amateur	Expert	Veteran
Upward	24.1%	42.3%	44.4%
Downward	75.9%	57.7%	55.6%

Table 2: Direction of the change.

Prompt	Stance	All	Amateur	Expert	Veteran
Analysis	Overweight	12.5%	11.8%	13.6%	6.6%
	Underweight	37.1%	50.0%	16.7%	7.6%
Promote	Overweight	23.7%	18.9%	31.8%	26.7%
	Underweight	40.4%	42.9%	36.4%	21.4%

Table 3: Influence of prompts and stances.

Stage	Amateur	Expert	Veteran
1st	61.2%	61.3%	62.2%
2nd	45.8%	44.7%	51.1%

Table 4: Accuracy of decisions.

sions for 50% of instances. To ensure the fairness of the experiment, we anonymized the stocks in all documents. This is intended to prevent participants from applying external knowledge, ensuring that their decisions are based solely on the information provided within the documents.

4 Behavioral Experiment

4.1 Preprocessing

The estimated cost of conducting experiments for all 234 instances is approximately 4,000 USD, which is prohibitively expensive. Therefore, we first adopt the Hierarchical Transformer-based Multi-task Learning model (HTML), utilized in financial forecasting based on ECCs (Yang et al., 2020), to simulate the experiment. To simulate the first stage of the experiment, we use additional neutral summaries from the ECTSum dataset to train the model. During the testing phase, we use the neutral summary and the analysis with stance as input to simulate the second stage. If the model's decision changes when given a summary and analysis, we select this summary-analysis pair for the human behavioral experiment. Ultimately, we have 75 instances for the experiment, reducing the cost to about 1,280 USD.[4]

4.2 Results and Analysis

Table 1 provides answers to RQ1 and RQ2. All experts have worked in the financial industry for more than five years, and we further group three experts with over ten years of experience as Veterans. First, the analysis written by professional analysts has a higher chance of changing experts' decisions. Second, amateurs are more likely to change their decisions based on GPT-4-generated analysis. Additionally, more experienced investors are less influenced by GPT-4-generated analysis. These results indicate that GPT-4's analysis may suffice for amateur scenarios but is still far from professional standards. It also echoes previous studies' concerns about human evaluation quality in natural language generation research (Snow et al., 2008; Howcroft et al., 2020), as many studies still evaluate models' outputs on crowdsourcing platforms. In other words, our results suggest that the analysis impacting amateurs may not be the focus for experts.

Table 2 further shows the direction of their decision changes. Upward (Downward) denotes a change in their predictions from degrease (increase) to increase (decrease). Overall, investors are more sensitive to underweight analysis, i.e., information that may negatively impact the company. However, the ratio between amateurs and experts is significantly different. This indicates that amateurs are very sensitive to negative information. This raises a potential risk of using LLMs to generate analysis for the general public. The generated underweight analysis has a higher potential to sway amateur investors' decisions, and our results provide evidence supporting the U.S. Department of Treasury's concerns about the risks of artificial intelligence in the financial services sector.[5] Imagining that automatically generated underweight analyses are widely distributed on online platforms, it may lead to higher market volatility and harm market stability.

To conduct an in-depth analysis of the risk, we further use GPT-4 to write promoting reports for the given stance. Table 3 shows the comparison. First, underweight analysis influences investors much more than overweight analysis. Second, analysis with a strong tone sways experts' decisions more than pure analysis, regardless of the given stance. This reveals the potential of LLMs in influencing professionals' decisions.

Finally, as mentioned in Section 4.1, we only

[4]More details about the setting of HTML are shown in Appendix B.

[5]https://home.treasury.gov/news/press-releases/jy2393

Annotator	Source	Grammatical	Convincing	Logical	Useful
Amateur	Analysis (GPT-4)	4.44	4.13	4.02	4.06
	Promote (GPT-4)	4.47	4.23	4.16	4.20
	Analyst	3.92	3.22	3.30	3.43
Expert	Analysis (GPT-4)	3.65	2.80	3.04	2.84
	Promote (GPT-4)	3.79	2.95	3.22	3.06
	Analyst	3.78	3.48	3.61	3.65
Veteran	Analysis (GPT-4)	3.71	2.78	3.03	2.46
	Promote (GPT-4)	3.79	2.95	3.22	3.06
	Analyst	4.06	3.93	4.09	3.97

Table 5: Multi-dimensional evaluation.

	Grammatical	Convincing	Logical	Useful
All	0.654	0.262	0.262	0.237
Amateur	0.505	0.109	0.136	0.179
Expert	0.769	0.317	0.391	0.169
Veteran	0.754	0.118	0.126	0.027

Table 6: Agreement among annotators.

focus on the pairs that lead the model to change decisions in spite of the accuracy. Thus, the analysis given in the second stage is not selected to lead investors to make wrong decisions. In Table 4, we show the accuracy of their decisions. The results reveal that investors make accurate trading decisions based on neutral summaries, and the analysis with stances may hurt the accuracy of their decisions. Based on this result, we want to highlight the risk of using generated analysis for financial decisions.

4.3 Generated Text Evaluation

Recently, many studies have scored generated text from multiple aspects (Zhong et al., 2022; Chan et al., 2023) to evaluate the quality of the generated documents. To answer (RQ3), we asked participants to annotate the given analysis from four aspects: grammar, convincingness, logical coherence, and usefulness. The score ranges from 1 to 5 (Discrete), with higher scores indicating better quality. Table 5 shows the average scores of different groups of participants for different sources.

First, from the objective aspect, i.e., grammar, GPT-4 achieves a level similar to that of professional analysts, regardless of the group of annotators. However, from the subjective aspects, amateurs and experts have different opinions on GPT-4-generated and analyst-written analyses. Amateurs provide higher scores for GPT-4-generated text, while experts provide higher scores for analyst-written analysis. These results highlight the difference between amateurs and experts. Given this evidence, future works should reconsider the design of the human annotation process.

Second, compared with the results in Section 4.2, experts change their decisions more frequently when analysts' reports are provided in the second stage, and these reports are considered more convincing, logical, and useful. The situation is similar for amateurs; GPT-4-generated analysis gets higher scores and leads to more changes in amateurs' decisions. This indicates that scores and reactions are correlated in our experiment. The correlation between scores and reactions in our experiment highlights the potential of using these reactions to evaluate forward-looking analyses, including predicting future stock trajectories with rationales. Finally, the experts' multi-dimensional evaluation scores also show the gap between state-of-the-art LLMs and professional analysts in writing analysis.

To check the agreement, each pair was annotated by at least two experts and two amateurs. We calculated Krippendorff's Alpha (Krippendorff, 2011), and the results are shown in Table 6. The agreement on grammatical scores is very high regardless of the annotators. This suggests that evaluating generated text from objective aspects is effective, as most studies did before the LLM era. However, the agreement on subjective metrics is quite low, even among experts. This indicates the problem of conducting human evaluation from subjective aspects, as different people have different opinions. Following the discussion of Amidei et al. (2018), the low agreement for complex generated text does not imply it is an insufficient evaluation metric, but it is natural after the generated text passes the Turing test. We hope the discussion in this paper can open different perspectives on generated text evaluation, particularly using readers' reactions as evaluation metrics.

5 Conclusion

This paper advocates for a nuanced approach to evaluating LLM-generated text and emphasizes the importance of real-world reactions as well as traditional evaluative metrics. By understanding and addressing the differences in how amateurs and experts perceive and are influenced by LLM-generated content, we can better harness the capabilities of these models while safeguarding against their potential pitfalls. Future research should continue exploring these dynamics, particularly focusing on the ethical implications and regulatory frameworks necessary to guide the responsible use of LLMs in decision-critical applications.

Limitations

First, the scope of our study is restricted to ECCs within the financial sector. Although this context is highly relevant for examining decision-making processes, the results may not be directly transferable to other domains where different types of information and decision-making criteria are involved. Future studies should explore a broader range of contexts to validate and expand upon our findings. Second, the sample size for our human behavioral experiment, though carefully selected to balance cost and representativeness, remains limited with 75 instances. This constraint may affect the statistical power and precision of our conclusions. Larger-scale studies are needed to confirm the trends and patterns observed in our research. Third, the evaluation of generated text involved subjective metrics such as convincingness, logical coherence, and usefulness, which inherently depend on individual perceptions. Despite efforts to mitigate this through multiple annotators and Krippendorff's Alpha calculation, the low agreement on subjective metrics highlights the challenge of achieving consistent evaluations across diverse groups. Developing more objective and standardized evaluation frameworks for LLM-generated text remains a critical area for future research.

Ethical Statements

This study deals with online experiments with a strong commitment to ethical standards in the treatment of participants. Prior to participation, all participants were provided with a comprehensive explanation of the study's objectives, the procedures involved, the potential risks, and their rights as study participants. Informed consent was obtained from all individual participants involved in the study. Participants were assured of their right to withdraw from the study at any point without any adverse consequences. To protect privacy, all data collected during the study were anonymized and securely stored. Identifiable information was removed from the dataset prior to analysis to ensure confidentiality. Participants were informed that the results of the study might be published, but privacy information would remain confidential and would not be linked to any personally identifying information. The online nature of the experiments was designed to ensure minimal risk to participants. However, appropriate measures were taken to address any technical and privacy-related concerns associated with online data collection.

References

Toyin D. Aguda, Suchetha Siddagangappa, Elena Kochkina, Simerjot Kaur, Dongsheng Wang, and Charese Smiley. 2024. Large language models as financial data annotators: A study on effectiveness and efficiency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10124–10145, Torino, Italia. ELRA and ICCL.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Maria Cristina Arcuri, Gino Gandolfi, and Ivan Russo. 2023. Does fake news impact stock returns? evidence from us and eu stock markets. *Journal of Economics and Business*, 125-126:106130.

Or Biran and Kathleen R McKeown. 2017. Human-centric justification of machine learning predictions. In *IJCAI*, volume 2017, pages 1461–1467.

Justin Birru, Sinan Gokkaya, Xi Liu, and René M Stulz. 2022. Are analyst short-term trade ideas valuable? *The Journal of Finance*, 77(3):1829–1875.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

Jonathan Clarke, Hailiang Chen, Ding Du, and Yu Jeffrey Hu. 2020. Fake news, investor attention, and market reaction. *Information Systems Research*, 32(1):35–52.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions.

In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Allen H Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Tero Karppi and Kate Crawford. 2016. Social media, financial algorithms and the hack crash. *Theory, culture & society*, 33(1):73–92.

Katherine Keith and Amanda Stent. 2019. Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 493–503, Florence, Italy. Association for Computational Linguistics.

Michael D Kimbrough. 2005. The effect of conference calls on analyst and market underreaction to earnings announcements. *The Accounting Review*, 80(1):189–219.

Shimon Kogan, Tobias J Moskowitz, and Marina Niessner. 2023. Social media and financial news manipulation. *Review of Finance*, 27(4):1229–1268.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1369–1385, New York, NY, USA. Association for Computing Machinery.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2022. ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Alexey Tikhonov and Ivan Yamshchikov. 2023. Post Turing: Mapping the landscape of LLM evaluation. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 398–412, Singapore. Association for Computational Linguistics.

Linyi Yang, Tin Lok James Ng, Barry Smyth, and Riuhai Dong. 2020. Html: Hierarchical transformer-based multi-task learning for volatility prediction. In *Proceedings of The Web Conference 2020*, pages 441–451.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Prompts for Text Generation with Investment Stance

Below are various prompts designed for text generation, each adopting a specific investment stance. The tasks are framed for different roles and stances:

Summarization: *As a financial analyst, you are tasked with preparing a detailed summary report on a recent earnings conference call transcript. Focus on key financial metrics.*
Transcripts: { }

Overweight analysis: *As a financial analyst, you are tasked with preparing a detailed summary report on a recent earnings conference call transcript, adopting an overweight investment stance. Focus on key financial metrics.*
Transcripts: { }

Underweight analysis: *As a financial analyst, you are tasked with preparing a detailed summary report on a recent earnings conference call transcript, adopting an underweight investment stance. Focus on key financial metrics.*
Transcripts: { }

Overweight Promotion: *As a stock promoter, you are tasked with preparing a report that offers a compelling promotion on the company, recommending an overweight investment stance based on the company's recent earnings call.*
Transcripts: { }

Underweight Promotion: *As a stock promoter, you are tasked with preparing a report that offers a cautious or skeptical perspective on the company,*

recommending an underweight investment stance based on the company's recent earnings call.
Transcripts: {}

B Details of HTML

We adopt different encoders with HTML, including BERT (Devlin et al., 2019), FinBERT-Tone (Huang et al., 2023), and FinBERT-Sentiment (Araci, 2019), and use Adam as the optimizer with an initial learning rate of 2e-5 (Yang et al., 2020). The model is trained for 10 epochs with a batch size of 4.