From Similarity to Consequences: Decision-Oriented Evaluation of Market Digest Generation

Yu-Shiang Huang,[1, 2] Chuan-Ju Wang,[1] Chung-Chi Chen[3]
[1]Academia Sinica, Taiwan
[2]National Taiwan University, Taiwan
[3]National Institute of Advanced Industrial Science and Technology, Japan
F09946004@ntu.edu.tw, cjwang@citi.sinica.edu.tw, c.c.chen@acm.org

Abstract

Natural language generation (NLG) is increasingly applied in finance, but its evaluation still relies on reference-based metrics that capture surface similarity rather than practical utility. This gap is critical for retail investors, who depend on short market digests, such as morning briefs and closing-bell reports, that have received little attention in prior research. Motivated by this gap, we define market digest generation as a new NLG task and benchmark it with performance-conditioned and professional-insight baselines. Beyond intrinsic metrics, we introduce a consequence-driven evaluation that measures how digests influence trading decisions made by both human investors and LLM agents. Our results show that LLM-generated morning briefs can improve decision accuracy over human references, while expert-curated asset selection further enhances outcomes, reaffirming the importance of human expertise. These findings underscore the limits of surface-level metrics and establish a decision-oriented framework for evaluating generated text by its real-world consequences.

1 Introduction

Natural language generation (NLG) systems are rapidly finding their way into high-stakes domains such as finance, medicine, and law, where the text they produce can influence decisions involving money, health, or even liberty. In finance, significant progress has been made in automating the generation of complex documents, such as financial reports and earnings call summaries, which primarily serve analysts and institutional clients.

However, a critical gap remains: these advances rarely translate into tools accessible to retail investors, who often rely instead on market digests delivered through easily accessible channels such as television, radio, or daily emails from trading apps. Such digests, typically morning briefs and closing-bell recaps, are short narrative summaries that condense international developments overnight or the day's key events, offering concise guidance in just a few minutes. Despite their importance for retail investors, market digest generation has received little direct attention in prior NLG research, and it presents unique challenges for evaluation.

Traditional intrinsic metrics, such as n-gram overlap scores (BLEU, ROUGE) or sentence-level plausibility, capture surface similarity but correlate only weakly with the actual utility of generated digests. A generated market digest may closely resemble human references, yet both can fail to provide real investment value. This highlights the danger of using similarity to human text as the primary criterion. Likewise, a fluent market brief may still mislead investors toward loss-making decisions. The central question, therefore, is not merely "how well" the generated text mirrors

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 Workshop on Generative AI in Finance.

reference phrasing, but "how effectively" it supports real-world decision-making in complex, dynamic environments.

In this work, we take the first step toward the systematic study of market digest generation, a practical yet underexplored scenario that directly serves retail investors. We formally define and benchmark this task and design several retrieval-augmented generation baselines. Most importantly, we propose a consequence-driven evaluation framework that goes beyond surface similarity and directly measures how both human-written and machine-generated digests influence investment decisions made by people and automated trading agents. By combining intrinsic metrics with decision-based assessment, our work establishes market digest generation as a distinct NLG task and highlights the critical need to evaluate generated text by its real-world consequences in high-stakes financial contexts.

2 Experimental Design

2.1 Human Reference Collection

We target two common forms of financial commentary: morning briefs (M_t) and closing-bell reports (C_t), both authored by professionals and used as human references. For each trading day t, we collect news articles N_t, stock price data P_t, listed entities E_t, and supplementary statistics S_t (e.g., market capitalization, institutional order flows, trading volumes). To build our dataset, we aligned 30 days of verbatim transcripts from financial news channels with their corresponding market events, yielding paired human baselines $(M_t^{\text{human}}, C_t^{\text{human}})$. On average, each day's commentary draws from $\sim$2,400 articles, with closing-bell reports additionally informed by intraday statistics. Dataset statistics are summarized in Table 5.

2.2 Market Digest Generation

We evaluate two classes of baselines for LLM-generated market digests, indexed by trading day t. Let M_t^{human} and C_t^{human} denote the human-authored morning and closing commentaries.

Performance-conditioned. To mimic how investors track extreme movers, we rank equities in E_{t-1} by volatility, trading volume, and institutional flows from (P_{t-1}, S_{t-1}), collect their articles N_{t-1}^{perf}, and generate

$$M_t^{\text{perf}} = \texttt{LLM}(N_{t-1}^{\text{perf}}).$$

For closing commentaries, a direct use of daily news N_t risks leakage since publication times do not necessarily reflect when events actually occurred, making it unclear which articles were available by the close.[1] We therefore condition the closing commentary on a morning brief, which is always produced before the closing bell, as a safe proxy:

$$C_t^{\text{perf}} = \texttt{LLM}(P_t, S_t, M_t^{\text{perf}}), \quad C_t^{M_{\text{human}}} = \texttt{LLM}(P_t, S_t, M_t^{\text{human}}).$$

These variants compare closing reports conditioned on generated versus expert-written morning narratives.

Professional-insight. From human commentaries, we extract entities mentioned in the morning brief ($E_{t-1}^{\text{pro-m}}$) and the closing report ($E_{t-1}^{\text{pro-c}}$), retrieve their news subsets, and generate

$$M_t^{\text{pro}} = \texttt{LLM}(N_{t-1}^{\text{pro-m}}), \quad C_t^{\text{pro}} = \texttt{LLM}(P_t, S_t, N_{t-1}^{\text{pro-c}}).$$

[2] This baseline evaluates the value of expert-curated entity selection and contrasts LLM versus human narrative quality under the same focus set.

Table 1 reports automatic scores against human references. Performance-conditioned morning briefs diverge sharply on ROUGE-Lsum, reflecting different entity choices from human writers. More broadly, reference-based metrics struggle to show which generations are actually useful. For instance, BertScore rates all baselines nearly the same. These limits highlight the need for decision-oriented evaluation, which we turn to next.

[1] By contrast, all prior-day news is certainly known by the next morning.

[2] Instead of feeding all articles in one batch, we generate snippets per entity and concatenate them, which keeps inputs manageable and guarantees coverage of all expert-highlighted companies.

Table 1: Automatic metrics for morning and closing scenarios. **SentM.** (Sentiment Match) is the proportion of cases where human and generated commentaries receive the same sentiment label (positive, negative, neutral) from an LLM-based classifier.

Scenario	Method	Len.	ROUGE-Lsum	BERTScore-F1	SentM.
Morning	M_t^{perf}	478.74	0.0479	0.6365	0.4831
	M_t^{pro}	12013.16	0.1790	0.6353	0.3820
Closing	C_t^{perf}	527.20	0.0975	0.6432	0.6067
	$C_t^{M_{\text{human}}}$	546.13	0.1086	0.6462	0.6404
	C_t^{pro}	8742.03	0.1396	0.6346	0.5169

2.3 Decision-oriented Evaluation

Not only do reference-based metrics provide little guidance on which digests carry greater value, but our qualitative observations also confirm this: LLMs generate fluent and coherent commentaries while diverging markedly from expert-written texts. Chasing lexical or semantic overlap is therefore uninformative. Instead, we treat generated texts as autonomous analytical inputs, to be judged by their ability to guide investment decisions.

To evaluate this decision value, we constrain scenarios to short horizons, minimizing confounds from long-term market trends. Participants—three human investors with different backgrounds[3] and three LLM agents (GPT-4o, Gemini-2.0-Flash, Claude-3.5-Sonnet)—used each digest to select only those stocks they were confident would move (by market close for morning briefs, or by the next open for closing reports). Importantly, participants were not required to forecast all assets but instead selected only those they felt confident enough to trade, reflecting realistic investor behavior.

Performance is measured by thresholded prediction accuracy, labeling returns above $+0.55\%$ as "rise" and below -0.50% as "fall" Xu and Cohen [2018]. These asymmetric thresholds reflect investors' greater aversion to losses and the need to ignore negligible price fluctuations that would not drive trading behavior. To ensure fairness, participants were prohibited from consulting external sources. Compensation included a base hourly wage set at 130% of the statutory minimum, plus bonuses: USD 65 for outperforming all LLMs, USD 100 for the top performer, and USD 35 for second place.[4]

3 Results and Analysis

Table 2 presents the evaluation results. Firstly, we observe significant differences in task difficulty between predicting intraday stock movements based on morning briefs and forecasting overnight market movements using closing-bell reports. The latter scenario appears inherently easier, as evidenced by consistently higher accuracy rates across all investors and information sources. Secondly, examining the morning brief scenario more closely, we find that decision accuracy based on original journalist-produced texts does not achieve optimal performance for either human or LLM investors. In contrast, LLM-generated morning briefs consistently lead to improved decision accuracy. This suggests that although traditional journalistic content is valuable, the analytical synthesis provided by LLM-generated summaries offers clearer signals or more actionable insights for immediate investment decisions.

Turning to the closing-bell reports, we observe a notable difference in utility between human and LLM investors. Journalist-produced closing-bell reports substantially benefit LLM agents, who achieve the highest accuracy when utilizing these professionally authored texts. However, the opposite trend emerges for human investors, who consistently perform better when decisions are informed by LLM-generated reports. This divergent outcome prompts reconsideration of traditional evaluation metrics: Should the ultimate goal be lexical and content alignment with human-generated expert texts, or should we instead assess the practical decision-making value that LLM-generated texts uniquely provide? Our experimental results strongly support the latter perspective, suggesting that LLM-generated content should be evaluated primarily for its practical efficacy rather than its fidelity

[3] The annotator background summary is provided in Table 4 in the appendix

[4] Annotation guidelines are provided in Appendix E.

Table 2: The accuracy of investor decisions under different market digests (%). The bolded results indicate the best performance of the same investor under identical decision-making scenarios. “Claude” refers to Claude-3.5-Sonnet and “Gemini” to Gemini-2.0-Flash.

		LLM Investor			Human Investor			
Scenario	Method	Claude	Gemini	GPT-4o	A	B	C	Average
Morning	M_t^{human}	38.85	44.89	42.35	37.97	36.67	34.40	39.19
	M_t^{perf}	45.98	**46.98**	42.53	42.92	**45.11**	**49.27**	**45.47**
	M_t^{pro}	**48.01**	42.41	**43.15**	**46.28**	40.23	48.35	44.74
Closing	C_t^{human}	**65.56**	**61.60**	**58.51**	48.13	50.83	42.24	54.48
	C_t^{perf}	49.42	49.57	43.29	51.32	45.11	65.06	50.63
	$C_t^{M_{\text{human}}}$	55.62	54.36	56.89	48.44	49.44	**75.00**	**56.63**
	C_t^{pro}	60.07	58.25	55.17	**54.05**	**53.45**	54.18	55.86
Investor’s Average		51.93	51.15	48.84	47.02	45.84	52.64	

to existing human-authored texts. Finally, regarding asset selection methods, the professional-insight approach generally enhances decision-making outcomes, reaffirming the critical role of human expertise in the content generation pipeline, with the only exceptions being the most experienced annotator and GPT-4o, whose performance was already comparable. Although exact replication of expert commentary may be unnecessary, human involvement in guiding the asset selection and content framing process significantly enhances decision quality. To further examine the characteristics of investment decisions, we provide a detailed behavior analysis in Appendix B.

4 Toward Decision-Oriented Evaluation

The variability observed in human investor performance presents significant challenges for reliably evaluating generated texts. Human decision-making is inherently influenced by personal biases, varying levels of domain expertise, and risk tolerance, which introduce substantial noise and complicate reproducibility and comparative analysis across different studies. Consequently, evaluating the efficacy of generated texts through human investors alone may yield inconsistent and difficult-to-generalize results.

In contrast, leveraging LLM investors offers distinct advantages for achieving consistent and reproducible evaluation outcomes. Under controlled settings (e.g., temperature set to zero), identical LLM models produce deterministic outputs, ensuring stable and replicable decision processes. By clearly documenting the model architectures, parameters, and input configurations, future research can precisely reproduce experimental conditions, thereby facilitating direct comparisons across studies and systematically assessing the impact of textual variations on decision-making efficacy.

Furthermore, this methodological framework extends beyond the use of contemporary LLMs for text evaluation, accommodating various text-based decision algorithms already developed within the computational finance and artificial intelligence communities. This flexibility allows for broader experimentation and benchmarking of different analytical approaches, enhancing our ability to rigorously evaluate and optimize textual content for decision support. In sum, adopting LLM investors as standardized evaluators can significantly enhance the reliability, reproducibility, and comparability of decision-oriented text evaluation, providing a robust foundation for future research.

Limitation

Our study has four main limitations. First, we focus on short-term decisions (intraday or overnight), which may not generalize to longer investment horizons that involve strategic planning and macroeconomic factors. Second, LLM investor agents, though deterministic under controlled settings, reflect training biases and lack the nuanced judgment of experienced professionals. Third, our analysis centers on financial digests (morning briefs and closing reports), limiting direct applicability to other domains such as medicine or law. Finally, we evaluate primarily with decision accuracy, which overlooks dimensions like risk-adjusted returns or portfolio diversity; future work should incorporate richer measures of decision quality.

Acknowledgments

This work was supported in part by AIST policy-based budget project "R&D on Generative AI Foundation Models for the Physical Domain."

References

Yen-Chun Hsu and Chenhao Tan. Decision-focused summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 611–626, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.47. URL https://aclanthology.org/2021.emnlp-main.47.

Shrey Joshi, Jyoti Singh, Saket Singh, and Sushant Tyagi. Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4567–4578. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.349. URL https://aclanthology.org/2022.findings-emnlp.349.

Takehiro Takayanagi, Hiroya Takamura, Kiyoshi Izumi, and Chung-Chi Chen. Can gpt-4 sway experts' investment decisions? In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 374–383, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.22. URL https://aclanthology.org/2025.findings-naacl.22.

Yumo Xu and Shay B. Cohen. Stock movement prediction from tweets and historical prices. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1183. URL https://aclanthology.org/P18-1183/.

Ningyu Zhang, Shumin Deng, Juan Li, Xi Chen, Wei Zhang, and Huajun Chen. Summarizing chinese medical answer with graph convolution networks and question-focused dual attention. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 15–24, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.2. URL https://aclanthology.org/2020.findings-emnlp.2.

A Related Work

Recent work in decision-oriented summarization emphasizes evaluating generated text based on its support for downstream decisions rather than surface-level quality. Hsu and Tan [2021] introduced metrics such as decision faithfulness and applied them to domains like healthcare. Takayanagi et al. [2025] showed that GPT-4-generated financial texts can influence investor behavior, with amateurs particularly susceptible to persuasive LLM-generated language. These studies underscore the importance of assessing real-world impact, yet they often focus on classification-style tasks or single-user decision accuracy, without closely examining how generated text affects actual downstream decision behavior in complex settings. In high-stakes domains such as finance and medicine, prior work highlights the dangers of misleading text and the necessity for outcome-based evaluation [Zhang et al., 2020, Joshi et al., 2022]. Building on these insights, our work introduces a consequence-driven evaluation framework tailored to financial market commentaries. Unlike prior studies, we systematically quantify how both human- and LLM-generated content affects concrete investment actions taken by human investors, LLM agents, and their collaboration for offering a more granular view of decision influence in practice.

B Investor Behavior Analysis

Table 3 further analyzes investor decision-making behavior by presenting the average number of transactions under various market digest scenarios. Across all conditions, both human and LLM investors consistently executed more buy than sell decisions. Notably, LLM investors engaged in a

Table 3: Average number of transactions.

Investor	Decision	Morning Briefs			Closing-Bell Reports		
		M_t^{human}	M_t^{perf}	M_t^{pro}	C_t^{human}	C_t^{perf}	C_t^{pro}
LLM	Buy	13.24	3.89	7.12	6.21	3.86	3.96
	Sell	4.17	1.19	1.30	2.37	2.06	1.28
Human	Buy	2.69	2.17	2.63	2.42	2.60	2.69
	Sell	1.36	0.56	0.96	1.22	1.33	1.01

higher overall number of transactions than human investors, especially when relying on journalist-produced market digests. This may be attributed to the nature of traditional financial journalism, which often references a wide array of assets and market signals without providing clear directional emphasis. Such ambiguity may prompt LLMs to interpret multiple weak or conflicting cues as actionable, resulting in a higher frequency of trades.

By contrast, LLM-generated market digests, particularly morning briefs, tend to present more distilled and prioritized information, which reduces signal ambiguity and discourages excessive trading. This effect is reflected in the overall reduction of transactions for both human and LLM investors when using LLM-generated content. In particular, the number of sell decisions declined notably, with human investors exhibiting a marked decrease in selling activity when guided by LLM-produced morning briefs. One possible explanation is that LLM-generated texts may be implicitly biased toward highlighting positive developments or investment opportunities. This tendency likely results from the model's training on general language data, which favors coherence, optimism, and recognizable patterns over uncertainty or negative sentiment. Consequently, this linguistic framing may influence both human and LLM investors to adopt a more optimistic outlook, increasing the likelihood of buy decisions while reducing sell-side activity.

C Background of Annotators

We asked the annotators to complete forms typically used by professional financial institutions to assess clients' investment risks. Table 4 presents the annotators' background information, including their professional experience, investment background, risk tolerance, and the size of their investable assets.

Table 4: Annotator background summary

Annotator	A	B	C
Working Industry	Wholesale & Retail	Information Technology	Financial Services
Investing Experience	2 years	4 years	7 years
Risk Aversion	Moderate	Moderate	Aggressive
Investing Budget	Low	Medium	High

D Dataset Statistics

Table 5 provides the statistics of the collected dataset.

E Annotation Guidelines

In this task, you will play the role of a Taiwan stock investor and make buy/sell decisions based on market news summaries. We hope to leverage your insights to capture subtle market shifts and compare your performance with that of large language models (LLMs).

Table 5: Summary Statistics of Human Commentaries and our dataset.

Statistics	M_t^{human}	C_t^{human}
Avg. Length	2,699.09	3,968.17
Avg. # of Entities	24.78	18.09
Avg. $\lvert N_{t-1} \rvert$	2,450.41	

E.1 Rewards and Incentives

You will be compensated at a base rate equal to 130% of the minimum hourly wage for each hour of annotation. If your decisions outperform the average performance of the participating LLMs, you will earn an additional USD 65. We have also prepared ranking bonuses among the three annotators: the first-place annotator will receive an extra USD 100, and the second-place annotator will receive an extra USD 35.

E.2 Annotation Workflow

1. **Read the Summary:** Carefully read one financial market news summary and fully understand the information.
2. **Adopt an Investor Mindset:** Imagine yourself as an active Taiwan stock investor, and consider what this summary implies for your portfolio.
3. **Make Your Decisions:** In the annotation interface, you will see two multi-select lists containing all Taiwan-listed stocks. You may select any number of stocks:
   - **Buy:** Select all stocks you predict will rise.
   - **Sell:** Select all stocks you predict will fall.
   - **No Action:** If the information is insufficient or you decide no trade is necessary, leave both lists empty and provide a brief remark (e.g., "Information too vague," "No directly relevant stocks").

   You are not limited to companies explicitly mentioned in the summary. Feel free to apply industry insights, economic trends, and your own investment experience and risk preferences to related firms.
4. **Remarks:** In the remarks field, provide a concise explanation of your overall portfolio strategy. Highlight only the most important points or the stocks you are most confident about, without detailing every selection.

E.3 Important Notes

- Do *not* consult real historical data or news articles—this would be considered cheating.
- Do *not* replicate actual past portfolios, as this would invalidate our evaluation.
- News summaries may be generated by humans or LLMs; do *not* let assumptions about the source influence your decisions.
- Multiple summaries may describe the same day's market open or close. Do *not* assume they refer to the same timestamp—make decisions based solely on the information presented.

E.4 Performance Evaluation

- **Open Summary:** Accuracy is measured against the same day's closing price.
- **Close Summary:** Accuracy is measured against the next day's opening price.
- A "Buy" decision is counted as correct if the stock rises by more than 0.55%.
- A "Sell" decision is counted as correct if the stock falls by more than 0.50%.
- All other outcomes are considered incorrect.
- Your success rate will be calculated by comparing your selections with actual stock movements.

Table 6: Illustrative examples of market commentary

Morning Brief	**Closing Bell**
Last Friday, U.S. markets closed higher—led by a 2.6% surge in the Philadelphia Semiconductor Index—after President Trump's weekend press conference omitted new China sanctions and trade-deal terminations, focusing instead on targeted visa bans and student restrictions. U.S. futures later pulled back as nationwide protests over a fatal police shooting in Minnesota rekindled racial tensions. Meanwhile, Huawei's August ban has spurred Chinese firms to accelerate chip self-production, Hong Kong's dollar remained freely convertible, and foreign investors sold NT$11 billion of Taiwan stocks. Today, investors await the U.S. ISM manufacturing index and Friday's non-farm payrolls, which may shape tomorrow's open . . .	At today's close, the Taiwan Stock Exchange rose 136 points to 11,079 on NT$166 billion in volume. Foreign investors returned, lifting electronics to 68% of turnover and driving the new market leader to NT$4,065 on strong volume. Analysts note that June's focus will be on companies with robust May revenues and the upcoming government voucher rollout boosting consumer sectors. Looking ahead, the electronics sector—anchored by TSMC and major chip designers—may sustain momentum if global reopening continues, while next week's earnings and policy announcements will guide the market's direction . . .

Please refer to the provided `companies.csv` file for a full list of listed stocks and their codes, which you can use in the annotation interface.

F Example

Table 6 presents examples of morning briefings and closing-bell commentaries.