LLM-as-a-Judge Evaluation of Financial News Articles generated based on Factors of Stock Price Fluctuation

Yurina Kosai, Yucheng Xie, Rikuto Tsuchida, Takehito Utsuro
Graduate School of Science and Technology, University of Tsukuba
1-1-1, Tennodai, Tsukuba, Ibaraki, 305-8573, Japan
{s2520763, s2620847, s2520796}_@_u.tsukuba.ac.jp, utsuro_@_iit.tsukuba.ac.jp

Abstract

This paper proposes an LLM-as-a-Judge evaluation framework of stock price fluctuation articles automatically generated based on financial news, corporate disclosures, and stock price fluctuation data. This automatic article generation framework emulates the workflow of human financial journalists by analyzing recent stock price fluctuations and incorporating relevant causal factors extracted from textual and numerical information. In particular, the generation process utilizes news articles and numerical stock price data, including price fluctuation ranges over the past three days. Based on those automatically generated stock price fluctuation articles, this study places particular emphasis on the LLM-as-a-Judge evaluation methodology. We conduct an item wise human evaluation and compare it with the LLM-as-a-Judge automatic metric. We analyze the correlation among these evaluation methods to assess their reliability. Furthermore, through comparisons between zero-shot and few-shot prompting, we examine the effectiveness of the proposed framework and the validity of LLM based evaluation for assessing factual and causal consistency in financial text generation.

Keywords: LLM-as-a-Judge, automatic evaluation, generating stock price fluctuation articles, factors of stock price fluctuation, large language models

1. Introduction

In providing information on stock price fluctuations, the usefulness of news articles extends beyond merely reporting the magnitude of price changes. Such articles also offer insights into the underlying factors that have led to those fluctuations. Typically, these articles are manually written for each individual stock. The conventional writing procedure is assumed to follow the format illustrated in Figure 1. Specifically, for stocks exhibiting large price fluctuations (either increases or decreases), journalists first identify information that may have contributed to the observed fluctuations. Subsequently, for stocks where such information is found, journalists summarize the relevant content and compose articles based on it. To address this task, automatic generation of stock price fluctuation explanation articles using large language models (LLMs) has recently attracted attention. Existing automatic generation methods (Nishida and Utsuro, 2025) achieve this by referring to numerical information on stock price fluctuations together with textual information that may serve as potential causal factors. In these approaches, official corporate IR disclosures are collected as textual information related to stock price fluctuation factors and used as the primary information source.

In contrast, this paper focuses on stock price fluctuation explanation article generation methods using LLMs (Nishida and Utsuro, 2025) and proposes an automatic evaluation method for the generated articles based on the LLM-as-a-Judge framework (Chiang and Lee, 2023; Zheng et al., 2023). Within the framework of this study, we first obtain stock price data for the most recent three days for each stock, based on the daily stock price change ranking published on the stock information website Kabutan[1], and use these data as factual information regarding stock price fluctuations[2]. Next, following (Nishida and Utsuro, 2025), we collect official corporate IR disclosures as textual information related to potential stock price fluctuation factors, and use these as the information source for generating stock price fluctuation explanation articles with an LLM. In addition, for these IR disclosures, we manually compose reference articles that explicitly clarify the relationship between the disclosed information and the stock price fluctuations. These serve as reference stock price fluctuation explanation articles. Based on these reference articles, we design a 10 point evaluation scale to assess content validity and clarity of explanation. Using this evaluation scale, we conduct both human evaluation of the automatically generated stock price fluctuation explanation articles and automatic evaluation under the LLM-as-a-Judge framework (Chiang and Lee, 2023; Zheng et al., 2023). In the evaluation experiments, we apply the LLM-as-a-Judge automatic evaluation scale to stock price fluctua-

[1] https://kabutan.jp/

[2] It should be noted that the article texts published on Kabutan are not used at all; only numerical data are utilized.

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 106–113

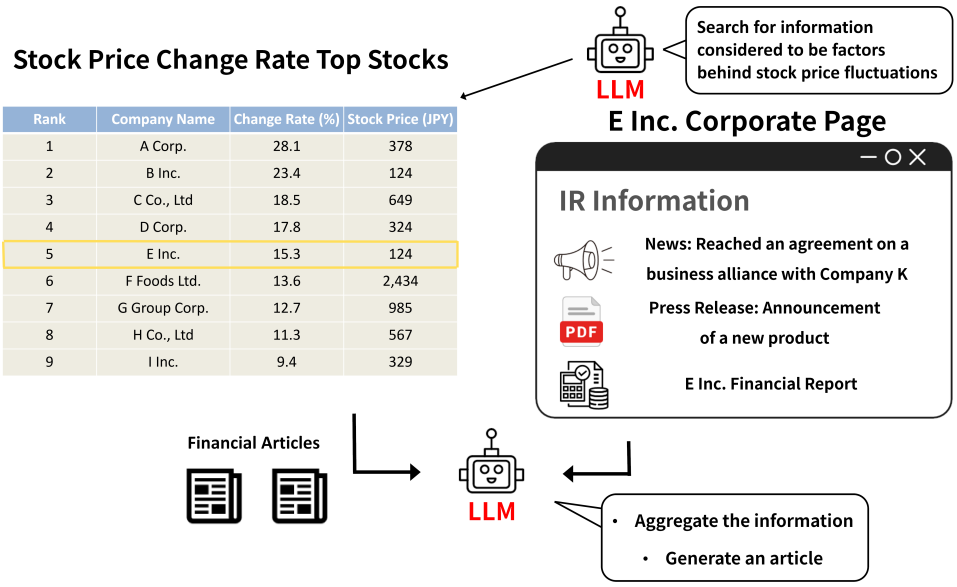


Figure 1: Generation of stock price fluctuation reason articles based on stock price fluctuation factor information (Nishida and Utsuro, 2025)

tion explanation articles generated by GPT-4o[3] in both zero-shot and few-shot settings, and achieve sufficiently high correlation with human evaluation results. In particular, higher correlation is observed for articles generated in the few-shot setting. In contrast, we show that the correlation between ROUGE scores (Lin, 2004) and human evaluation results is extremely low, thereby demonstrating the effectiveness of the proposed automatic evaluation scale under the LLM-as-a-Judge framework (Chiang and Lee, 2023; Zheng et al., 2023).
The main contributions of this paper are summarized as follows:

- We propose an LLM based automatic evaluation framework for stock price fluctuation article generation, grounded in the LLM-as-a-Judge paradigm, to assess the semantic adequacy and explanatory quality of generated financial articles.
- We construct a benchmark setting that combines numerical stock price data and official corporate IR disclosures, while manually creating reference articles to ensure clear causal alignment between price fluctuations and explanatory content.
- We design a task specific 10 point evaluation scheme that decomposes stock price fluctuation articles into key components (stock price information, explanation of fluctuation factors, and supplementary information), allowing fine-grained comparison between human and automatic evaluations.
- Through experiments on GPT-4o generated articles (zero-shot and few-shot), we demonstrate that the proposed LLM based evaluation method achieves a strong correlation with human judgments, substantially outperforming ROUGE in this task.

2. Related Work

Regarding the LLM-as-a-Judge, which employs LLMs as evaluators, it has been reported that automatic evaluation by LLMs can serve as a viable alternative to human evaluation (Chiang and Lee, 2023; Zheng et al., 2023). For example, Zheng et al. (2023) demonstrated the potential of utilizing LLMs for assessing the subjective quality of generated responses. Furthermore, Liu et al. (2023) proposed a framework in which evaluation criteria are explicitly provided to the LLM, which then generate a chain-of-thought and perform step-by-step scoring in a form-based input format. Their results show that evaluation using GPT-4 substantially outperforms conventional automatic metrics in terms of correlation with human judgments. Meanwhile, Saha et al. (2024) proposed a branch-solve-merge approach, in which evaluation criteria are decomposed into multiple aspects, each of which is assessed individually, and the results are subsequently aggregated. They report that this method achieved higher agreement with human evaluation compared to single-pass holistic evaluation. In Imajo et al. (2025), a reference answer set based evaluation framework is constructed along three dimensions — fluency, truthfulness, and helpfulness — and its consistency with LLM based evaluation results is demonstrated.
In the research activities targeting question an-

[3] https://openai.com/ja-JP/api/

swering (QA) tasks where concrete reference answers to the questions do exist, Badshah and Sajjad (2025) proposed a majority voting evaluation method using multiple LLMs. They showed that combining multiple models improves evaluation reliability and achieved strong correlation with human evaluation. On the other hand, Bai et al. (2023) introduced stepwise scoring and ranking based on multiple criteria, including accuracy, coherence, factuality, and comprehensiveness, and reported that the resulting evaluation outcomes exhibit high agreement with human annotations.

3. Stock Price Fluctuation Factor Information

In this study, we use officially disclosed corporate information as potential factors underlying stock price fluctuations. Specifically, we collect IR disclosures and timely disclosure documents released by companies through the stock price exchanges on which they are listed, and treat them as stock price fluctuation factor information. These disclosures include content that may affect stock prices, such as financial results announcements, revisions of earnings forecasts, changes in business strategies, announcements of new products or services, and business alliances.

4. Target Stock Price Fluctuation Articles for Analysis

We focus on stock price fluctuation articles that contain both numerical stock price information and explanations of the price fluctuations based on corporate IR disclosures. Concretely, we utilize stock price data from the "ranking today" section provided by the stock information website Kabutan. This ranking targets stocks with the highest daily percentage increases and decreases, thereby enabling daily identification of stocks exhibiting significant price fluctuations.

In this paper, we target articles that include both stock price information and explanations of stock price fluctuations derived from the corresponding companies' IR disclosures. By aligning these articles with stock price fluctuation factor information (see the next section), we perform automatic generation of stock price fluctuation explanation articles.

5. Alignment between Stock Price Fluctuation Factors and Stock Price Fluctuation Articles

Given a stock price fluctuation article, for the designated company of the stock price fluctuation article, we collect corporate official disclosures concerning that company, then extract information that can be regarded as plausible factors explaining the observed stock price fluctuation. Specifically, we target corporate communications that were released prior to the publication date of the stock price fluctuation article, and establish correspondences based on semantic relatedness. In this alignment process, we allow not only exact lexical matches but also paraphrased or summarized expressions to be considered as corresponding information.

6. Article Generation from Stock Price Fluctuation Factors

In this study, we generate stock price fluctuation explanation articles using a large language model, taking as input stock price fluctuation factor information extracted from financial news and corporate disclosures, and recent numerical stock price data. This section describes the task formulation, input construction, prompt design, and generation strategy (Nishida and Utsuro, 2025).

The generation task in this study is formulated as a conditional text generation problem, where the model outputs a natural language stock price fluctuation explanation article conditioned on stock price fluctuation factor information and numerical stock price data. The input consists of the company name, numerical information including the direction and magnitude of the stock price fluctuation, and factor sentences that have been identified as causally related to the stock price fluctuation. The output is a short article written in a style comparable to that written by human financial journalists, concisely describing both the stock price fluctuation and its underlying causes. For numerical stock price information, we provide the closing prices over the most recent trading days and the magnitude of fluctuation, ensuring that the direction of the stock price fluctuation is clearly specified. For stock price fluctuation factor information, we use only those sentences extracted from financial news and corporate IR disclosures that are determined to have a causal relationship with the stock price fluctuation. This design prevents the inclusion of general industry descriptions or background information that lack direct causal relevance, thereby improving the precision of explanation generation.

In generating stock price fluctuation explanation articles, we employ GPT-4o as the LLM. The prompt explicitly instructs the model to act as a financial market journalist, to restrict the content to the provided input information, to avoid fabricating numerical values, to clearly state the direction of the stock price fluctuation, and to concisely describe the reasons for the change. In the zero-shot setting, only the task description and input information are

provided. In the few-shot setting, manually written example articles are included to guide stylistic and structural consistency. The temperature is set to 0 during generation to increase output determinism and ensure reproducibility in evaluation.

Under this framework, numerical data and textual factor information are integrated to generate explanations grounded in causal relationships. In particular, by explicitly providing stock price fluctuation factor information as input, the proposed method enables the generation of financial market articles with enhanced explainability and factual consistency.

7. Evaluation of Stock Price Fluctuation Articles

7.1. Overview

This section describes the human evaluation and the LLMs based automatic evaluation metrics employed to assess the stock price fluctuation articles generated by the LLM. In this study, ROUGE (Lin, 2004) is adopted as a baseline automatic evaluation metric. ROUGE is a lexical overlap based metric that measures similarity to reference texts; however, it is insufficient for evaluating the validity of explanations, such as whether the generated article appropriately captures the causes of stock price fluctuations. In particular, because stock price fluctuation articles typically contain a relatively small proportion of stock related terms within the entire text, ROUGE, which relies on lexical overlap, tends to produce unstable evaluations and is not necessarily well suited to this task. Therefore, this study proposes an evaluation metric based on LLMs, which has recently attracted considerable attention. We further analyze its correlation with human evaluation to demonstrate the effectiveness of automatic evaluation methods for stock price fluctuation article generation.

7.2. Manually Developing Reference Articles

As reference articles, 40 stock price fluctuation articles were manually written with reference to articles published on Kabutan between December 8 and 11, 2025. Each manually written stock price fluctuation article includes numerical information on stock prices, the primary cause of the price fluctuation, and supplementary explanations regarding the company or its products when necessary. The length of each article ranges from 150 to 350 Japanese characters, maintaining a level of conciseness comparable to that of actual stock price fluctuation articles. A concrete example is shown in Table 1. In describing the reasons for stock price fluctuations, we referred to the companies' disclosed IR information and based the descriptions on content judged to be directly related to the observed price fluctuations.

7.3. Evaluation Criteria

This section describes the evaluation criteria used to assess the quality of the generated stock price fluctuation articles. A stock price fluctuation article generally consists of multiple components, including (i) stock price information, (ii) an explanation of the factors underlying the price fluctuation, and (iii) supplementary information regarding the company or its business. Accordingly, we designed an evaluation scheme that separately assesses each of these components.

Specifically, human evaluation was conducted on a 10 point scale, divided into the following three categories:

- **Stock Price Information (4 points)**

  This criterion evaluates whether expressions related to stock price fluctuations, such as price increases or decreases, are correctly described and consistent with the actual price fluctuation.

- **Explanation of Stock Price Fluctuation Factors (3 points)**

  This criterion assesses whether the underlying factors behind the stock price fluctuation are accurately explained based on the company's disclosed information.

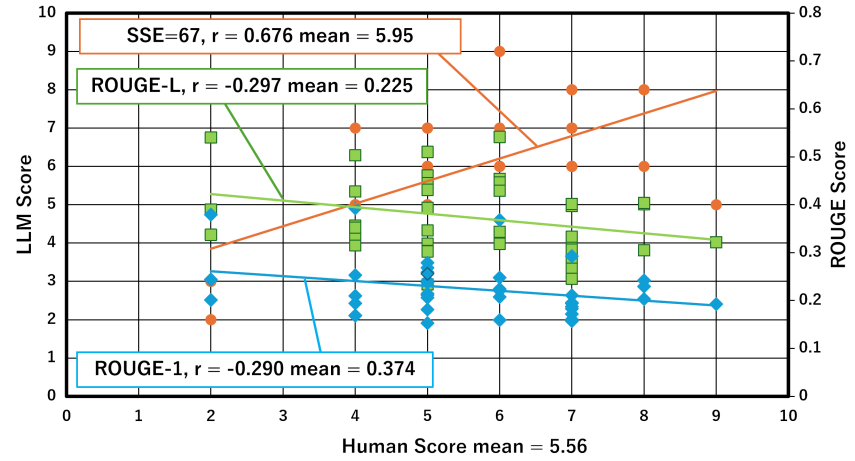
- **Additional Information (3 points)**

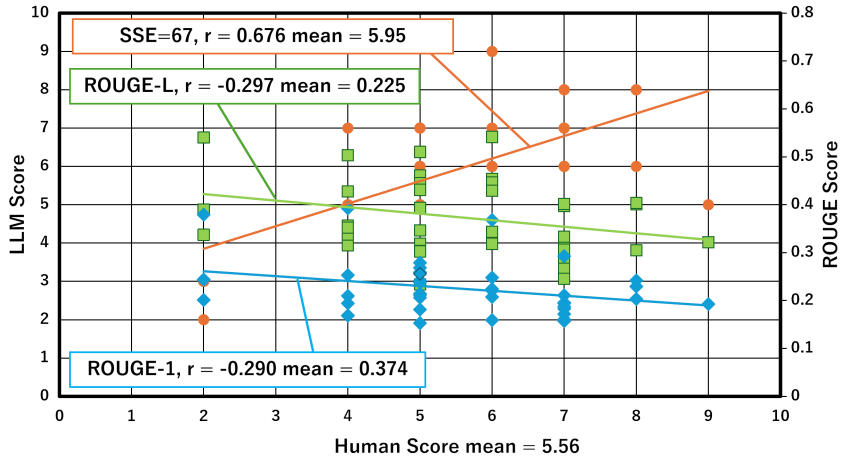
  Points are assigned based on either of the following types of information. The scores for these two types are not cumulative; instead, points are awarded only for the type that contains more substantial information: (i) a detailed explanation of the stock price fluctuation factors, or (ii) the accuracy of supplementary information, such as future outlooks or related business activities.

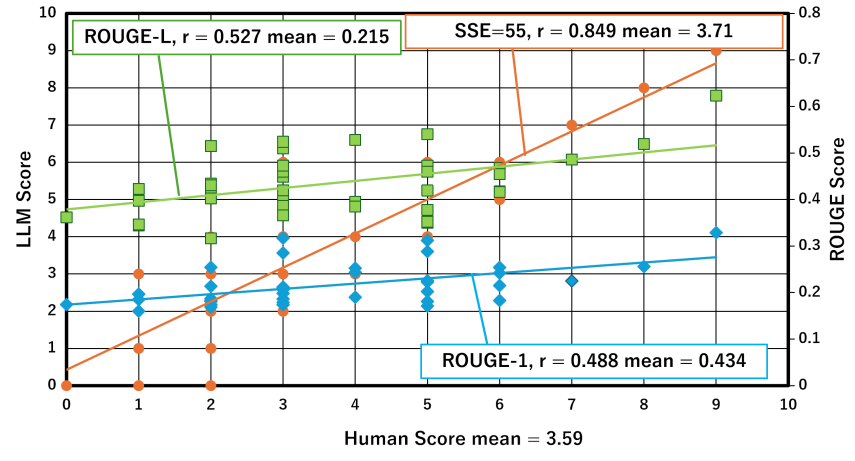
The total evaluation score for each article is calculated as the sum of the points assigned to each category. By separating the evaluation criteria in this manner, we aim to analyze not merely the textual similarity, but the extent to which the essential elements required in a stock price fluctuation article are satisfied.

7.4. Correlation Analysis between Human and Automatic Evaluation Results

In this section, we first generated 40 stock price fluctuation articles using GPT-4o (zero-shot and 5-







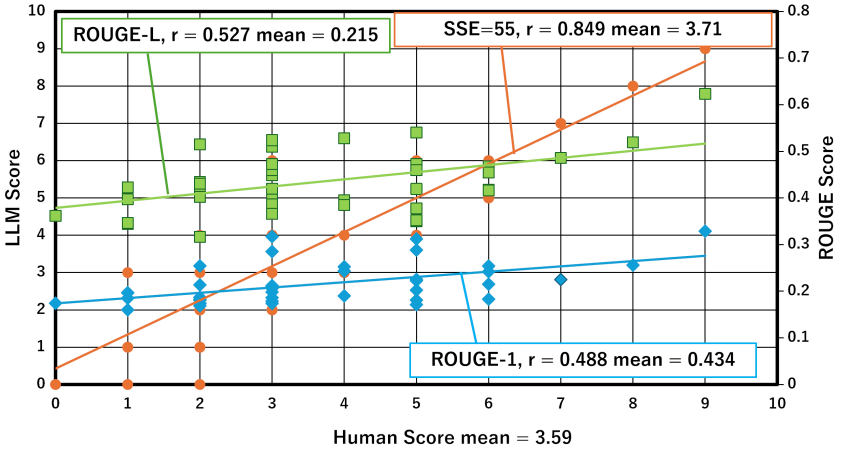


Figure 2: Correlation analysis results with human evaluation scores: Comparison of article evaluation results by LLM (Gemini 3, zero-shot) and ROUGE-1/L (F1 score).

Manually Written	Generated by GPT-4o (zero-shot)	Generated by GPT-4o (5-shot)
BrightPath Biotherapeutics <4594> continued its sharp rise. After the close of trading on the 8th, the company announced that it had received a notice of patent approval from the Japan Patent Office regarding a patent application related to its development pipeline BP1209 (fully personalized neoantigen vaccine). The patent concerns linker technology used in BP1209, a fully personalized cancer treatment neoantigen vaccine. This news was viewed as a positive catalyst.	The stock price of BrightPath <4594> rose from 50 yen to 52 yen and then to 55 yen over three days. The increase was attributed to the company's announcement that it had received a notice of patent approval from the Japan Patent Office regarding its development project BP1209. The patent relates to a fully personalized neoantigen vaccine for cancer treatment, and the technological progress was positively evaluated by the market.	The stock price of BrightPath Biotherapeutics <4594> is rising. The company announced that a patent related to BP1209 had received a notice of patent approval from the Japan Patent Office. This patent acquisition concerning BP1209, a fully personalized neoantigen vaccine for cancer treatment, is considered to have raised future expectations and supported the stock price.
Number of characters: 160	Number of characters: 163	Number of characters: 140
	Human: Stock price: 2 points Reason: 2 points Others: 1 point Total: 5 points	Human: Stock price: 2 points Reason: 2 points Others: 2 points Total: 6 points
	LLM: Stock price: 2 points Reason: 3 points Others: 1 point Total: 6 points	LLM: Stock price: 3 points Reason: 2 points Others: 1 point Total: 6 points
	ROUGE-1: 0.539, ROUGE-L: 0.458	ROUGE-1: 0.589, ROUGE-L: 0.521

Table 1: Example of a manually written article and articles generated by LLM (GPT-4o, zero-shot / 5-shot) (Article text, number of characters. Human and LLM (Gemini 3, zero-shot) evaluation results (Stock price: 4 points; Reason for fluctuation: 3 points; Others: 3 points; Total: 10 points), ROUGE-1/L (F1 score))

shot settings) and conducted a correlation analysis between human evaluation results and automatic evaluation results.

For automatic evaluation, we employed the evaluation criteria described in the previous section and used Gemini 3 provided by Google[4] as a zero-shot evaluator. By adopting the LLM-as-a-judge framework, the evaluator can comprehensively assess semantic consistency and logical coherence between the generated text and the reference text. Compared with ROUGE, which is based on n-gram overlap, this approach is more suitable for evaluating textual quality that involves structural and explanatory elements.

To measure the relationship between human and automatic evaluation results, we used the correlation coefficient r and the sum of squared errors (SSE). Let y_i and $\hat{y}_i$ $(i = 1, \ldots, n)$ denote the human evaluation score and the automatic evaluation score, respectively. SSE is defined as the sum of squared differences between the predicted values obtained from LLM based evaluation and the observed values obtained from human evaluation, as follows:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

A smaller SSE indicates a smaller discrepancy between the automatic (LLM based) evaluation and the human evaluation.

The correlation analysis results for the 40 reference articles are shown in Figure 2.

For the articles automatically generated by GPT-4o in the zero-shot setting, the LLM based automatic evaluation showed a strong positive correlation with human evaluation ($r = 0.676$), and the discrepancy from human scores was relatively small (SSE = 67). In contrast, ROUGE-1 ($r = -0.290$) and ROUGE-L ($r = -0.297$) exhibited negative correlations, indicating low consistency with human evaluation results.

For the articles generated by GPT-4o in the 5-shot setting, the correlation between automatic evaluation and human evaluation further improved ($r = 0.849$), demonstrating a very strong positive correlation. The discrepancy from human evaluation was also reduced compared to the zero-shot setting (SSE = 55). However, ROUGE-1 ($r = 0.527$) and ROUGE-L ($r = 0.488$) showed only moderate or lower levels of correlation.

These results indicate that, in the stock price fluctuation article generation task, the proposed method of using an LLM as an evaluator achieves a high correlation with human evaluation tendencies. In contrast, ROUGE was found to be inappropriate as an evaluation metric for stock price fluctuation articles.

[4] https://gemini.google/jp/about/?hl=ja

7.5. Case Study

This section presents concrete examples of stock price fluctuation articles automatically generated by GPT-4o in the zero-shot and 5-shot settings (Table 1) and analyzes the differences among human evaluation, LLM based evaluation, and ROUGE.

In the zero-shot setting, the generated article described specific numerical transitions in the stock price (50 yen $\rightarrow$ 52 yen $\rightarrow$ 55 yen), thereby supplementing numerical details. However, since the manually created reference article did not include explicit numerical transitions, this addition was judged as unnecessary information. As a result, the human evaluation assigned 2 points for stock price information, 2 points for the explanation of the price fluctuation factors, and 1 point for additional information, for a total of 5 points. Although the stock price expression was described merely as "increase," whereas the reference article used the term "continued rise," the direction of the fluctuation was consistent; therefore, 2 points were awarded. For the explanation of the price fluctuation factors, the reproduction of key terms such as "patent approval," "BP1209," and "fully personalized neoantigen vaccine" was positively evaluated. In comparison, the LLM based evaluation largely reproduced the human evaluation results, although it differed in assigning 3 points for stock price information.

In contrast, in the few-shot setting, the evaluation of stock price information and the explanation of fluctuation factors was the same as in the zero-shot case, but the appropriateness of supplementary information improved. Specifically, unnecessary numerical supplementation was suppressed, and the information was organized more coherently in accordance with the context. As a result, the human evaluation assigned 2 points for stock price information, 2 points for the explanation of fluctuation factors, and 2 points for additional information, for a total of 6 points. In the LLM based evaluation, differences were observed in two categories—3 points for stock price information and 1 point for additional information—yet the total score was consistent with the human evaluation.

8. Conclusion

In this paper, we proposed an automatic evaluation method for stock price fluctuation article generation using LLMs, based on the LLM-as-a-Judge framework (Chiang and Lee, 2023; Zheng et al., 2023), targeting the LLM based stock price fluctuation article generation method proposed in (Nishida and Utsuro, 2025). Specifically, the generated stock price fluctuation articles were evaluated automatically by an LLM under an item wise evaluation scheme.

In the evaluation experiments, we applied the proposed LLM-as-a-Judge-based automatic evaluation metric to stock price fluctuation articles automatically generated by GPT-4o in zero-shot and few-shot settings. The results demonstrated a sufficiently high correlation with human evaluation scores, indicating the effectiveness of the proposed automatic evaluation approach.

As future work, it will be necessary to establish an LLM based automatic evaluation method for stock price fluctuation articles generated from non-textual information sources, such as numerical data and chart data.

9. Bibliographical References

Sher Badshah and Hassan Sajjad. 2025. Reference-guided verdict: LLMs-as-judges in automatic evaluation of free-form QA. In *Proceedings of the 9th Widening NLP Workshop*, pages 251–267.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2023. Benchmarking foundation models with language-model-as-an-examiner. In *Proceedings of the 37th Advances in neural information processing systems*, pages 78142–78167.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.

Kentaro Imajo, Masanori Hirano, Shuji Suzuki, and Hiroaki Mikami. 2025. A judge-free LLM open-ended generation benchmark based on the distributional hypothesis. *arXiv preprint arXiv:2502.09316*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-EVAL: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

Shunsuke Nishida and Takehito Utsuro. 2025. Headline generation for stock price fluctuation articles. In *Proceedings of the 6th Workshop on Financial Technology and Natural Language*

Processing and Multi-Lingual ESG Impact Type Identification Shared Task, pages 184–195.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2024. Branch-solve-merge improves large language model evaluation and generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8352–8370.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proceedings of the 37th Advances in neural information processing systems*, pages 46595–46623.