

Preliminary Study on LLM-as-a-Judge Evaluation of Causal Mechanism Explanations in Financial News and its Bias Analysis

Yucheng Xie, Yurina Kosai, Rikuto Tsuchida, Takehito Utsuro

Graduate School of Science and Technology, University of Tsukuba

1-1-1, Tennodai, Tsukuba, Ibaraki, 305-8573, Japan

{s2620847, s2520763, s2520796}@_u.tsukuba.ac.jp, utsuro_@_iit.tsukuba.ac.jp

Abstract

Evaluating causal mechanism explanations in generated financial news is challenging because stock price movements may have multiple plausible explanations without a unique gold explanation. We propose a reference-free LLM-as-a-Judge framework that evaluates explanations from three perspectives: Relevance, Completeness, and Economic Plausibility. We construct 200 articles from Chinese stock market data and compare Gemini 3.5 Flash, GPT-5.5, and Claude Opus 5 with human annotations under zero-shot and few-shot settings. Before aggregating the evaluation results, we applied a systematic matched-subset procedure to examine potential self-preference and generation-source bias. Based on the data examined in this study, these analyses found no statistically significant evidence of systematic bias, supporting the aggregation of all 200 articles for the overall agreement analysis. The results show moderate to high agreement with human evaluation, with Completeness generally showing higher agreement than Relevance. Few-shot prompting does not consistently improve agreement and generally leads to more conservative judgments. These findings demonstrate the potential of LLM-as-a-Judge for reference-free evaluation of causal mechanism explanations in financial news.

1 Introduction

Large Language Models (LLMs) have recently shown promising performance in financial text generation and automatic evaluation. Previous studies have demonstrated the feasibility of generating stock price fluctuation articles and evaluating them using LLM-as-a-Judge on the Japanese stock market (Nishida and Utsuro, 2025; Kosai et al., 2026). However, existing work mainly evaluates overall article quality or explanations based on company-related information.

For investors, understanding why a stock price fluctuates is more important than writing quality

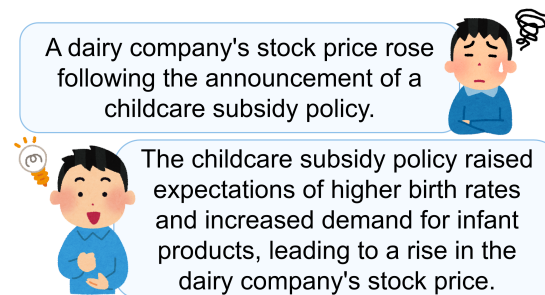


Figure 1: Example of causal mechanism explanations for a stock price movement.

alone. This is particularly relevant to the Chinese stock market, where stock price movements can be influenced by government policies, industry trends, and investor sentiment. Policy signals, in particular, have been identified as an important factor affecting investors' decisions in China (Hou and Yang, 2021). As illustrated in Figure 1, simply stating that a policy announcement was followed by a stock price increase does not explain why the two are related. An appropriate causal mechanism explanation should describe the intermediate effects through which the policy may influence the company's business and, ultimately, its stock price. Therefore, evaluating the quality of causal mechanism explanations for stock price movements is an important but challenging problem.

In this paper, we propose an LLM-as-a-Judge framework specialized for evaluating causal mechanism explanations in stock price fluctuation articles. Given an observed event and the corresponding stock price movement, the generated article describes a causal mechanism that connects the two, and our task is to evaluate the quality of this proposed mechanism explanation. Because the actual mechanisms underlying individual stock price movements are generally unknown, our goal is not to establish causality itself, but to evaluate the quality of the causal mechanism explanations proposed

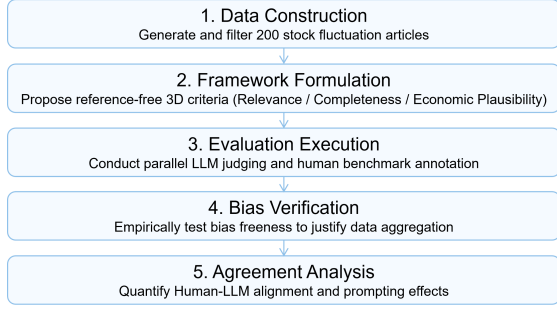


Figure 2: Overview of the pipeline.

in generated articles. We evaluate explanations from three perspectives: **Relevance**, **Completeness**, and **Economic Plausibility**. As summarized in Figure 2, we first construct and filter 200 stock fluctuation articles and formulate the reference-free evaluation framework, followed by parallel LLM judging and human annotation. Because the articles are generated using different models and generation methods, these generation sources could potentially affect LLM judgments. Therefore, before aggregating the evaluation results, we use matched subsets to systematically examine generation-model, generation-method, and self-generation effects and assess whether evaluation results from different generation sources can be reasonably combined. Based on the data examined in this study, we found no statistically significant evidence of such bias, allowing the 200 articles to be analyzed together for the subsequent agreement analysis.

The main contributions of this paper are:

- We propose an LLM-as-a-Judge framework for evaluating causal mechanism explanations in stock price fluctuation articles.
- We introduce a three-dimensional evaluation framework consisting of Relevance, Completeness, and Economic Plausibility.
- We examine potential evaluation bias associated with the generation model, generation method, and self-generation using matched subsets, providing a systematic procedure for assessing such bias.
- We compare multiple LLM judges and evaluation methods in terms of their agreement with human annotations.

2 Related Work

Previous work proposed an LLM-based framework for generating stock price fluctuation articles from

financial information and stock price data (Nishida and Utsuro, 2025). We build on this framework to generate articles for evaluation, focusing on causal mechanism explanations of stock price movements.

LLM-as-a-Judge has been widely studied for evaluating generated text. Recent surveys have summarized its applications, reliability challenges, and approaches to mitigating evaluation biases (Gu et al., 2026). Li et al. (Li et al., 2025) further provided a systematic taxonomy of LLM-as-a-Judge along evaluation targets, judging methods, and benchmarking approaches. Zheng et al. (Zheng et al., 2023) demonstrated the potential of LLMs for assessing subjective qualities of generated responses. G-Eval (Liu et al., 2023) further showed that reference-free LLM evaluation can achieve substantial correlation with human judgments, while raising concerns about bias toward LLM-generated texts. Thakur et al. (Thakur et al., 2025) further showed that even strong LLM judges can differ from human evaluations and exhibit vulnerabilities to prompt complexity and leniency bias. Shi et al. (Shi et al., 2025) identified position bias in LLM judges and showed that its strength varies across judges and tasks and is influenced by the quality gap between evaluated responses. Chen et al. (Chen et al., 2024) showed that both human and LLM judges can exhibit judgment biases, highlighting the need to examine the reliability of evaluation results. These findings motivate our analysis of whether evaluation results are affected by the generation model and generation method. Wataoka et al. (Wataoka et al., 2024) found that LLM judges favored lower-perplexity outputs regardless of whether they were self-generated, suggesting that self-preference may reflect a broader preference for familiar outputs rather than whether the outputs were self-generated. Koo et al. (Koo et al., 2024) also reported substantial cognitive biases in LLM evaluators, including egocentric bias toward their own outputs. Chen et al. (Chen et al., 2025) noted that directly comparing scores on self-generated and other-model-generated responses may confound self-preference with response quality differences. Following this principle, we control for response quality by comparing judge-human residuals between self-generated and non-self-generated articles.

Chain-of-Thought (CoT) prompting (Wei et al., 2022; Kojima et al., 2022) elicits intermediate reasoning steps and has been shown to improve complex reasoning. We use CoT as a generation method

and examine whether it affects subsequent LLM-as-a-Judge evaluation.

LLMs have also demonstrated financial knowledge and practical abilities. Son et al. (Son et al., 2023) investigated financial reasoning in LLMs and showed that their ability to generate coherent financial reasoning emerges with sufficient model scale. Wang et al. (Wang et al., 2024) used LLMs to extract factors influencing stock movements from financial news and provide explanations for stock movement prediction. FinEval (Guo et al., 2025) introduced a benchmark for financial knowledge and practical tasks in the Chinese financial domain, showing that current LLMs possess a certain level of financial capability. Nie et al. (Nie et al., 2025) further proposed a comprehensive Chinese financial benchmark covering financial knowledge, qualifications, practical tasks, and financial law, highlighting both the capabilities and remaining challenges of LLMs in the financial domain. This provides a basis for investigating their use in evaluating financial explanations.

LLM-as-a-Judge has also been applied to stock price fluctuation articles (Kosai et al., 2026). Their framework compares generated articles with corresponding reference articles, adopting a reference-based setting. In contrast, our framework does not require a reference explanation, as multiple valid causal mechanism explanations may exist for the same stock price movement.

Koupae et al. (Koupae et al., 2025) showed that LLMs still struggle to accurately identify causal connections between events and proposed causal event graphs to explicitly represent causal chains for event reasoning.

Evaluating explanations is challenging because a given event may admit multiple valid reasoning paths. Jansen et al. (Jansen et al., 2021) studied explanation evaluation in multi-hop science question answering, where explanations are constructed from facts retrieved from a knowledge base. They proposed evaluating explanations from the perspectives of Relevance and Completeness: whether individual facts are relevant to the question and answer, and whether the selected facts together form a sufficient reasoning chain. Following these concepts, our framework evaluates whether generated explanations contain relevant statements and sufficiently cover necessary causal information, while additionally introducing Economic Plausibility for financial explanations.

3 Underlying LLM-as-a-Judge Framework of Evaluating Causal Mechanism Explanations

This section presents the proposed framework for evaluating LLM-generated stock price fluctuation articles. Each article is evaluated from three perspectives: **Relevance** (1–5), **Completeness** (1–5), and **Economic Plausibility** (0, 1, or Not Evaluable (NE)).

The framework adapts the Relevance and Completeness criteria of Jansen et al. (Jansen et al., 2021) to stock price fluctuation explanations and introduces Economic Plausibility to assess the financial reasonableness of the inferred causal relationships.

Relevance (1–5). The original Relevance criterion evaluates whether each supporting fact is relevant to the target question and answer. Unlike explanation selection tasks, generated financial news articles are not constrained by a fixed knowledge base, and LLMs may introduce unsupported causal statements that are not grounded in the provided inputs. Therefore, we adapt Relevance to evaluate whether each generated statement is justified by the available information and appropriate to include in the explanation. Thus, Relevance measures whether a statement should appear in the article, with unsupported or irrelevant statements receiving lower scores.

Completeness (1–5). Completeness evaluates whether the article covers the important information and reasoning steps necessary to explain the stock price movement. Higher scores indicate a sufficiently complete causal chain connecting the given event to the stock price movement, including important intermediate factors such as business effects, investor expectations, and market sentiment. Omitted information or causal steps result in lower scores. Since it is difficult to define a universally complete explanation without a reference explanation, judges assess completeness based on the important information and causal steps they expect from the given event and stock price movement. The validity of this judgment is then examined through its agreement with human evaluation, with higher correlation indicating greater consistency with human judgments of explanatory completeness.

Economic Plausibility (0, 1, NE). While Relevance and Completeness evaluate whether an explanation is supported by the input information and

sufficiently covers the reasoning chain, they do not assess whether the inferred causal relationship is economically reasonable. Therefore, we introduce Economic Plausibility to evaluate whether the proposed causal mechanism explanation is consistent with general financial knowledge.

Economic Plausibility uses three labels: **1** (economically plausible), **0** (economically implausible), and **Not Evaluable (NE)**. NE is assigned when the article lacks sufficient information to determine whether the proposed causal relationship is economically reasonable, typically because the causal chain or supporting evidence is incomplete. NE therefore does not indicate an incorrect explanation. Forcing such cases into either plausible or implausible would require unsupported assumptions and introduce additional evaluation noise.

Relevance and Completeness are scored on a scale from 1 to 5, with 1 indicating the lowest and 5 the highest level of performance. Detailed definitions of each score, together with the complete evaluation prompts, are provided in the Appendix.

Figure 3 illustrates the three evaluation dimensions and their roles in assessing causal mechanism explanations.

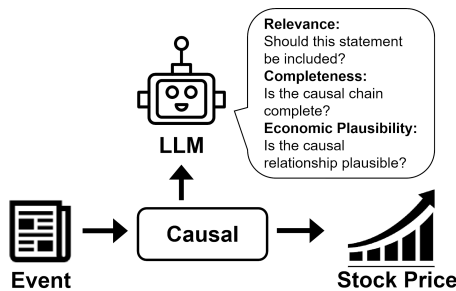


Figure 3: Illustration of the three evaluation dimensions: Relevance, Completeness, and Economic Plausibility.

4 Experimental Evaluation Procedure

This section describes the dataset, experimental setup, and evaluation metrics used in this study.

4.1 Dataset

We constructed a dataset of stock price fluctuation articles based on events in the Chinese stock market during April 2026. We first collected 277 event-level instances, each containing the date, company information, stock price movement, and corresponding financial event. Company information was collected and organized by an LLM using a search engine. The financial events and corresponding stock price movements were collected

from CLS¹, a Chinese website that publishes real-time financial and stock market news. The original CLS event descriptions were not directly used as input information; instead, all collected events were manually rewritten before being provided as input for article generation. For each instance, articles were generated using Gemini 3.5 Flash² and GPT-5.5³ with two generation methods, Direct and Chain-of-Thought (CoT), resulting in 1,108 articles. The prompts used for article generation are provided in the Appendix.

The generated articles were then screened using an LLM-based screening procedure that assessed the relevance and completeness of the generated explanations, and samples covering different apparent quality levels were selected based on these assessments. The selected articles were manually reviewed, and those containing clearly incorrect event or company information were removed. This resulted in 200 articles for the final evaluation dataset, as shown in Table 1.

Table 1: Distribution of the final dataset.

Generation Model / Method	Articles
Gemini 3.5 Flash / Direct	99
Gemini 3.5 Flash / CoT	52
GPT-5.5 / Direct	26
GPT-5.5 / CoT	23
Total	200

4.2 Experimental Setup

Human evaluation was conducted by a single student annotator with some knowledge of finance. To reduce the potential influence of limited financial expertise, the annotator consulted external materials when necessary to clarify general financial concepts and terminology. However, such searches were restricted to general financial knowledge and did not include information about the specific companies, events, or stock price movements being evaluated. All human judgments were therefore based on the same input information provided to the LLM judges, while the supplementary searches were used only to clarify general financial knowledge.

Three LLMs, Gemini 3.5 Flash, GPT-5.5, and Claude Opus 5⁴, were evaluated as automatic judges. Each judge evaluated the generated arti-

¹<https://www.cls.cn/>

²<https://aistudio.google.com/>

³<https://openai.com/api/>

⁴<https://platform.claude.com/>

cles according to the proposed framework. Gemini 3.5 Flash was run with a temperature of 0 and default settings for all other API parameters, while GPT-5.5 and Claude Opus 5 were run with their default API settings, as temperature is not available as a user-configurable parameter for these models.

Two prompting strategies were compared: zero-shot and few-shot. In the zero-shot setting, only the evaluation instructions and scoring criteria were provided. In the few-shot setting, the same prompt was supplemented with ten manually selected examples covering different score levels and explanation qualities. Each example included the input information, generated article, human scores, and scoring rationale.

For each article, the judge assigned scores for Relevance, Completeness, and Economic Plausibility, together with textual explanations. These judgments were analyzed using the methods described in the following subsection.

4.3 Evaluation Scheme

For Relevance and Completeness, we calculated Pearson’s correlation coefficient (r) and Spearman’s rank correlation coefficient (ρ).

Economic Plausibility was treated as a three-class classification. Cohen’s κ was therefore used to measure agreement between LLM and human judgments.

We further examined whether judges exhibit systematic evaluation biases and whether their judgments are affected by the generation source of the evaluated articles. The significance of these effects was assessed through statistical comparisons of evaluation tendencies, generation methods, and generation models.

5 Empirical Study on Evaluation Bias

To examine whether the generation source of an article affects LLM-as-a-Judge evaluation, we developed a systematic procedure based on paired analyses under controlled generation conditions. All LLM judges were evaluated under the zero-shot setting in this analysis. For Relevance and Completeness, we used their harmonic mean as a summary measure, hereafter referred to as the F-score for convenience. Relevance measures whether the article focuses on information that effectively explains the stock price movement and is supported by the provided input, whereas Completeness measures whether the explanation covers the key causal

steps connecting the event to the observed stock price movement. These two criteria therefore capture complementary aspects of explanation quality: an explanation may be highly relevant but incomplete, or complete but contain substantial irrelevant or unsupported content. Their complementary roles are analogous to those of precision and recall, for which the harmonic mean is commonly used to obtain a balanced summary that is sensitive to low values in either dimension. The F-score was calculated as

$$F = \frac{2 \times \text{Relevance} \times \text{Completeness}}{\text{Relevance} + \text{Completeness}}.$$

This aggregation was also necessary because, under some generation conditions, Pearson correlations were unstable or unavailable due to limited score variation; in particular, Gemini assigned the highest Completeness score to all CoT-generated articles in the first comparison described below. Thus, this F-score serves as a summary measure specific to our evaluation framework rather than the conventional classification F-score.

We considered two comparisons with relatively sufficient sample sizes: (1) Gemini-generated articles with Direct versus CoT generation ($n = 32$), where the generation model, input information, and stock-price event were fixed and only the generation method varied; and (2) Direct articles generated by Gemini 3.5 Flash versus GPT-5.5 ($n = 13$), where the generation method, input information, and stock-price event were fixed and only the generation model varied. The corresponding comparisons for GPT-5.5 between Direct and CoT and for CoT articles between Gemini 3.5 Flash and GPT-5.5 each contained only four matched events and were therefore not included in the statistical comparisons.

Given the limited sample sizes, these findings should be interpreted as conclusions supported by the data examined in this study, with their generalizability requiring further investigation.

For Table 2, r_D and r_{CoT} denote the Pearson correlations between judge and human F-scores under Direct and CoT generation, respectively; EP_D and EP_{CoT} denote the corresponding agreement rates with human EP labels. The Steiger ZPF test was used to compare the dependent F-score correlations, while McNemar’s test was used to compare the paired EP agreement rates. As shown in Table 2, the F-score correlations decreased from Direct to CoT for all three judges. However, none of the

Table 2: Effect of generation method on F-score correlation and EP agreement ($n = 32$).

Judge	r_D	r_{CoT}	ZPF p	EP_D	EP_{CoT}	McNemar p
Claude	0.599	0.297	0.132	0.469	0.812	0.007
GPT	0.670	0.306	0.058	0.781	0.781	1.000
Gemini	0.624	0.454	0.354	0.719	0.875	0.227

Table 3: Effect of generation model on F-score correlation and EP agreement ($n = 13$).

Judge	r_{Gem}	r_{GPT}	ZPF p	EP_{Gem}	EP_{GPT}	McNemar p
Claude	0.644	0.427	0.473	0.308	0.615	0.219
GPT	0.384	0.338	0.900	0.769	0.846	1.000
Gemini	0.436	0.603	0.600	0.692	0.846	0.625

Table 4: Self-preference bias on matched Gemini- and GPT-generated articles ($n = 13$).

Judge	$\overline{Res}_{self}$	$\overline{Res}_{nonself}$	Δ_F	Wilcoxon p	r_{self}	$r_{nonself}$	ZPF p	EP_{self}	$EP_{nonself}$	McNemar p
Gemini	-0.256	+0.418	-0.674	0.110	0.436	0.603	0.600	0.692	0.846	0.625
GPT	-0.365	-0.145	-0.220	0.542	0.338	0.384	0.900	0.846	0.769	1.000

differences was statistically significant under the Steiger ZPF test. For EP, Claude showed a significant increase in exact agreement ($0.469 \rightarrow 0.812$, McNemar’s $p = 0.007$), whereas the changes for GPT and Gemini were not significant. Further inspection revealed that Claude assigned NE to 15 of 32 articles under Direct generation, but to none under CoT generation. This may reflect Claude’s more conservative evaluation tendency under the Direct setting, particularly in cases where the economic plausibility of an explanation was difficult to determine. Consequently, the apparent improvement in EP agreement after introducing CoT was largely driven by the reduction of NE assignments rather than a clear improvement in distinguishing $EP=0$ and $EP=1$. Among the cases that became agreements under CoT, 10 of 13 (77%) were cases in which Claude had assigned NE under Direct while the human annotation was $EP=0$ or $EP=1$. After excluding these NE-related cases, the agreement increased only from 0.765 to 0.824, and the difference was no longer significant (McNemar’s $p = 1.000$).

In Table 3, r_{Gem} and r_{GPT} represent the Pearson correlations between the judge’s F-scores and the human F-scores for articles generated by Gemini and GPT, respectively. The Steiger ZPF p value tests whether the two correlation coefficients differ significantly. EP_{Gem} and EP_{GPT} denote the proportions of cases in which the judge’s EP label agrees with the human label for Gemini- and GPT-generated articles, respectively, and McNemar’s p value tests whether the paired agreement rates differ significantly. As shown in Table 3, no statistically significant difference was found in either F-score correlation or EP agreement. Although some numerical differences were observed, none of the Steiger ZPF or McNemar tests was signifi-

cant. Thus, within the matched sample, changing the generation model from Gemini to GPT did not provide clear evidence of a systematic change in LLM-as-a-Judge evaluation behavior.

We additionally examined whether judges systematically favored articles generated by themselves. Because the generated articles may differ in their underlying quality, we controlled for response quality by calculating the F-score residual for each article as the difference between the judge’s F-score and the human F-score:

$$Res_J(x) = J^F(x) - H^F(x).$$

For each judge, we calculated the mean residual for self-generated and non-self-generated articles, denoted by $\overline{Res}_{self}$ and $\overline{Res}_{nonself}$, respectively. We then defined

$$\Delta_F = \overline{Res}_{self} - \overline{Res}_{nonself}.$$

A positive Δ_F indicates that the judge assigned relatively higher F-scores to self-generated articles. The paired residual differences were tested using the Wilcoxon signed-rank test.

In Table 4, r_{self} and $r_{nonself}$ denote the Pearson correlations between judge and human F-scores for self-generated and non-self-generated articles, respectively, with the ZPF p value testing whether the two dependent correlations differ significantly. EP_{self} and $EP_{nonself}$ denote the exact agreement rates between the judge and human EP labels for self-generated and non-self-generated articles, respectively, with McNemar’s p value testing whether the paired agreement rates differ significantly.

Within our experimental setup and matched dataset, neither Gemini nor GPT showed statistically significant self-preference bias in the F-score residual, Pearson correlation, or EP agreement analyses. Gemini showed a negative Δ_F of -0.674 ,

while GPT showed a smaller negative difference of -0.220 , but neither difference was statistically significant according to the Wilcoxon signed-rank test. The Pearson correlations with human scores also did not differ significantly between self-generated and non-self-generated articles according to the Steiger ZPF test. Likewise, neither judge showed a significant difference in EP agreement under McNemar’s test. Thus, within the available matched sample, we found no statistically significant evidence that either judge systematically favored articles generated by itself.

6 Correlation with Human Evaluation

The preceding bias analysis, based on the data examined in this study, found no statistically significant evidence that the generation method or generation model systematically affected LLM-as-a-Judge evaluation, including self-preference bias. Therefore, within our dataset, we treat the evaluation scores as sufficiently independent of the generation source and combine all 200 generated articles for the following agreement analysis. Agreement results for each of the four model-prompting combinations are reported separately in Appendix D. Although the bias analysis used a harmonic mean of Relevance and Completeness as a summary score for the paired comparisons, the following agreement analysis considers Relevance and Completeness separately rather than aggregating them into an overall score. Table 5 summarizes the average scores assigned by the human annotator and the three LLM judges, while Table 6 reports their agreement. All reported correlation coefficients were statistically significant at $p < 0.001$ under both zero-shot and few-shot settings, indicating strong statistical evidence for the observed associations. The complete confusion matrices for the three evaluation dimensions are provided in Appendix E.

Overall, all three LLM judges showed moderate to high agreement with human evaluation, although the agreement varied across dimensions. **Completeness** consistently showed higher agreement than **Relevance**. Across both prompting settings, Pearson’s r ranged from 0.557 to 0.664 for Relevance, compared with 0.679 to 0.775 for Completeness. This suggests that judging whether an explanation contains sufficient causal information is more consistent than determining whether individual statements should be included in the expla-

nation.

For **Relevance**, Claude Opus 5 achieved the highest agreement under zero-shot prompting ($r = 0.664$, $\rho = 0.693$). For **Completeness**, the highest agreement was achieved by Claude Opus 5 under zero-shot prompting ($r = 0.773$, $\rho = 0.787$) and by Gemini 3.5 Flash under few-shot prompting ($r = 0.775$, $\rho = 0.784$).

Claude Opus 5 generally showed the strongest agreement on the two score-based dimensions, particularly under zero-shot prompting.

For **Economic Plausibility**, GPT-5.5 achieved the highest agreement under zero-shot prompting ($\kappa = 0.591$), followed by Gemini 3.5 Flash ($\kappa = 0.522$) and Claude Opus 5 ($\kappa = 0.326$). The larger variation among judges suggests that EP is more sensitive to the characteristics of the evaluation model.

The average scores in Table 5 show a general tendency for LLM judges to assign lower scores than the human annotator. The human mean scores were 4.145 for Relevance and 3.770 for Completeness, while the LLM means ranged from 3.265 to 4.125 and from 3.005 to 3.915, respectively. This indicates that LLM judges generally adopted a more conservative scoring standard, particularly for Relevance.

A closer examination of the generated articles suggests that the larger gap in Relevance may be partly because LLM judges tend to regard some background information as unnecessary content, whereas the human annotator was more tolerant of such background information.

For EP, human annotations contained 85.0% EP=1, 11.5% EP=0, and 3.5% NE. GPT-5.5 showed a similar distribution under zero-shot prompting, with 84.0% EP=1, 14.5% EP=0, and 1.5% NE, consistent with its relatively high agreement. In contrast, Claude Opus 5 assigned NE to 24.5% of the samples, consistent with the tendency observed in Section 5 for Claude to assign NE more frequently to Direct-generated articles. This may indicate a relatively conservative evaluation tendency when economic plausibility could not be clearly determined.

We also examined agreement among the three LLM judges to assess the consistency of their evaluations across models, as shown in Table 7. Notably, Spearman’s ρ was consistently higher for Completeness (0.812~0.872) than for Relevance (0.657~0.774), mirroring the pattern observed in human-LLM agreement. This suggests that the

Table 5: Average evaluation scores. Rel and Comp denote the mean Relevance and Completeness scores, respectively. EP₀, EP₁, and EP_{NE} denote the percentages of the corresponding EP labels.

Evaluator	Prompting	Rel	Comp	EP ₀ (%)	EP ₁ (%)	EP _{NE} (%)
Human	–	4.145	3.770	11.5	85.0	3.5
Claude Opus 5	Zero-shot	3.310	3.115	17.0	58.5	24.5
	Few-shot	3.265	3.005	11.5	56.0	32.5
Gemini 3.5 Flash	Zero-shot	4.125	3.915	13.0	82.5	4.5
	Few-shot	3.635	3.415	22.5	61.5	16.0
GPT-5.5	Zero-shot	3.825	3.330	14.5	84.0	1.5
	Few-shot	3.690	3.160	14.0	73.0	13.0

Table 6: Agreement with human evaluation. Pearson’s correlation (r), Spearman’s correlation (ρ), and Cohen’s κ for Economic Plausibility (EP).

Model	Prompting	Rel (r)	Rel (ρ)	Comp (r)	Comp (ρ)	EP (κ)
Claude Opus 5	Zero-shot	0.664	0.693	0.773	0.787	0.326
	Few-shot	0.625	0.651	0.726	0.738	0.259
Gemini 3.5 Flash	Zero-shot	0.596	0.619	0.718	0.733	0.522
	Few-shot	0.584	0.618	0.775	0.784	0.282
GPT-5.5	Zero-shot	0.557	0.540	0.705	0.716	0.591
	Few-shot	0.557	0.548	0.679	0.689	0.401

Table 7: Agreement among LLM judges. Spearman’s ρ is used for Relevance and Completeness, and Cohen’s κ for Economic Plausibility.

Judge Pair	Rel (ρ)	Comp (ρ)	EP (κ)
Claude–GPT	0.734	0.855	0.302
Claude–Gemini	0.774	0.872	0.380
GPT–Gemini	0.657	0.812	0.530

higher consistency of Completeness is not specific to comparisons with human evaluation, but is also observed across different LLM judges.

Overall, Completeness was the most consistently aligned with human evaluation and also showed the highest agreement among LLM judges, while Relevance was comparatively more difficult to judge. Economic Plausibility showed greater variation across evaluation models and among LLM judges, suggesting that this dimension is more sensitive to the choice of judge.

7 Effect of Few-shot Prompting

Table 5 and Table 6 compare the evaluation results under zero-shot and few-shot prompting.

Few-shot prompting did not consistently improve agreement with human evaluation. To test whether the differences were statistically significant, we applied Steiger’s test for dependent overlapping correlations to Pearson’s r and Spearman’s ρ , and paired bootstrap resampling as a robustness check. For EP, we tested the difference in Cohen’s

κ using paired bootstrap resampling.

For **Relevance**, none of the three judges showed a significant difference between zero-shot and few-shot prompting ($p > 0.05$). For **Completeness**, agreement significantly decreased for Claude Opus 5 (r : 0.773 $\rightarrow$ 0.726, $p = 0.046$; ρ : 0.787 $\rightarrow$ 0.738, $p = 0.025$) but significantly increased for Gemini 3.5 Flash (r : 0.718 $\rightarrow$ 0.775, $p = 0.022$; ρ : 0.733 $\rightarrow$ 0.784, $p = 0.031$). No significant change was observed for GPT-5.5.

For **Economic Plausibility**, agreement significantly decreased for all three judges. The changes in κ were -0.067 (95% CI $[-0.134, -0.003]$) for Claude Opus 5, -0.190 (95% CI $[-0.318, -0.061]$) for GPT-5.5, and -0.240 (95% CI $[-0.368, -0.100]$) for Gemini 3.5 Flash.

A notable effect of few-shot prompting was a shift toward more conservative judgments. For all three judges, both Relevance and Completeness scores decreased, while the proportion of NE judgments for EP increased. This suggests that the demonstrations may have provided the judges with more concrete patterns for identifying weaknesses and assigning lower scores or NE labels.

The improvement in Gemini 3.5 Flash’s Completeness evaluation appears to be related to this effect. Under zero-shot prompting, Gemini frequently assigned the maximum Completeness score, whereas the few-shot examples provided more patterns for identifying missing causal infor-

mation and assigning lower scores. This tendency may also be influenced by the seemingly complete causal chains produced under the CoT generation setting, which could make the explanations appear more complete to the judge. This resulted in a score distribution that was closer to the human annotations and consequently improved its agreement.

Overall, few-shot prompting primarily changed the evaluation tendency rather than consistently improving agreement with human evaluation.

8 Case Study

Table 8 presents a case in which the four evaluators assigned different scores to the same generated article. The company name is anonymized in the generated article, while external entities mentioned in the input event are retained. All LLM judges were evaluated under the zero-shot setting.

Table 8: Case study of different evaluation tendencies.

Section	Content (Translated from Chinese)																				
Background	Date: 2026/4/14 Company: Anonymized company Recent five-day stock prices: 27.42, 26.95, 27.28, 27.64, 29.43 Event: Bloom Energy, a U.S. fuel-cell company, announced on Monday that it had expanded its partnership with Oracle. Oracle would purchase a total of 2.8 GW of Bloom Energy’s fuel-cell systems. Following the announcement, Bloom Energy’s stock price rose by more than 15% in after-hours trading and reached a new historical high. Company information: The company specializes in the investment, construction, and operation of long-distance natural gas pipelines. Its business also includes gas storage and peak-shaving facilities and distributed energy projects. Its gas supply area covers approximately 70% of the province.																				
	As of April 14, 2026, the company, which specializes in the investment, construction, and operation of long-distance natural gas pipelines, showed an upward trend with fluctuations. Its stock prices over the past five days were 27.42, 26.95, 27.28, 27.64, and 29.43, with a significant increase on the final day. The stock price movement was mainly driven by a positive development in the international energy sector. On Monday, U.S. fuel-cell company Bloom Energy announced an expanded partnership with Oracle, under which Oracle would purchase a total of 2.8 GW of Bloom Energy’s fuel-cell systems. Following the announcement, Bloom Energy’s stock price rose by more than 15% in after-hours trading and reached a new historical high. Since the company is involved not only in natural gas pipeline construction but also in gas storage, peak-shaving facilities, and distributed energy projects, this positive development in the energy sector naturally generated a positive spillover effect on the company. As a company whose gas supply area covers approximately 70% of the province, the company was boosted by the event, and its stock price ultimately rose to 29.43 on the fifth day.																				
Generated Article																					
Scores	<table><tr><th>Evaluator</th><th>Rel</th><th>Comp</th><th>EP</th></tr><tr><td>Human</td><td>4</td><td>3</td><td>1</td></tr><tr><td>Claude</td><td>3</td><td>2</td><td>NE</td></tr><tr><td>GPT</td><td>4</td><td>3</td><td>1</td></tr><tr><td>Gemini</td><td>4</td><td>4</td><td>1</td></tr></table>	Evaluator	Rel	Comp	EP	Human	4	3	1	Claude	3	2	NE	GPT	4	3	1	Gemini	4	4	1
	Evaluator	Rel	Comp	EP																	
	Human	4	3	1																	
	Claude	3	2	NE																	
	GPT	4	3	1																	
Gemini	4	4	1																		

The case illustrates differences in evaluation tendencies among the evaluators. Human and GPT assign identical scores, considering the article sufficiently relevant but only moderately complete. Gemini gives a higher Completeness score because

it regards the causal chain from the Bloom Energy–Oracle event to the company’s stock price as largely connected. In contrast, Claude assigns lower Relevance and Completeness scores, emphasizing that the article contains considerable repetition of the input information and does not sufficiently explain how the fuel-cell event is transmitted to the company’s natural gas and distributed-energy businesses or to investor expectations.

The Economic Plausibility scores also illustrate a difference in judgment criteria. Human, GPT, and Gemini regard the proposed causal relationship as economically plausible, while Claude assigns NE because the available information is insufficient to determine whether the connection between fuel-cell demand and the company’s natural gas pipeline business is economically justified. Thus, the NE label does not necessarily indicate an implausible explanation, but can reflect insufficient evidence for evaluating its economic plausibility.

9 Conclusion

We proposed a reference-free LLM-as-a-Judge framework for evaluating causal mechanism explanations in stock price fluctuation articles. The framework evaluates explanations from three perspectives: Relevance, Completeness, and Economic Plausibility, with an NE label for cases where Economic Plausibility cannot be reliably judged from the available information.

Our matched-subset procedure found no statistically significant evidence of systematic self-preference or generation-source effects in the examined samples. Based on these findings, we combined the generated articles for the overall agreement analysis. Experiments on 200 generated articles showed moderate to high agreement between LLM-based and human evaluations. Completeness generally achieved higher agreement than Relevance, while Economic Plausibility showed larger differences across judges. Few-shot prompting did not consistently improve agreement and generally made the judges more conservative, although it improved Gemini’s Completeness evaluation by reducing its tendency to assign the highest score.

These results provide preliminary evidence that LLM-as-a-Judge can be useful for reference-free evaluation of causal mechanism explanations in financial news, while also indicating that its evaluation behavior can vary depending on the evaluation dimension and prompting strategy.

Limitations

There are two main limitations to the present study. First, the dataset is relatively small, particularly for the matched subsets used in the paired analyses. Although statistically significant results were observed in several comparisons, the limited sample size reduces the statistical power of the analyses and makes it difficult to detect smaller effects. A larger dataset would therefore be necessary to obtain more robust statistical evidence and to examine whether the observed tendencies generalize to a broader range of stock price fluctuation events.

Second, the human evaluation was conducted by a single annotator who was a student with some knowledge of finance. Although the annotator consulted relevant materials during the evaluation process to reduce ambiguity and improve consistency, relying on a single annotator limits the reliability of the human evaluation and prevents direct assessment of inter-annotator agreement. In particular, judgments of Relevance, Completeness, and Economic Plausibility may involve some degree of subjectivity, even when the evaluation criteria are clearly defined.

Third, the few-shot prompting setting was not systematically varied in this study. We used a fixed set of demonstration examples and did not examine how factors such as the number of examples, their score distributions, or the characteristics of the demonstrated evaluations affect LLM-as-a-Judge performance. Therefore, the observed effects of few-shot prompting may depend on the particular demonstrations used.

Future work should therefore expand the dataset and recruit multiple human annotators with stronger or more diverse financial expertise. Inter-annotator agreement could then be measured to assess the reliability of the human evaluation, and disagreements could be used to further refine the evaluation criteria and prompting instructions. In addition, a larger and more diverse set of stock price events and generated articles would allow the proposed framework to be evaluated under a wider range of generation and evaluation conditions. Future work should also systematically vary the number and composition of few-shot demonstrations, including their score distributions and other characteristics, to investigate how different few-shot settings influence evaluation consistency and agreement with human judgments.

References

- Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. [Humans or LLMs as the judge? a study on judgement bias](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.
- Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. 2025. [Beyond the surface: Measuring self-preference in LLM judgments](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1653–1672, Suzhou, China. Association for Computational Linguistics.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2026. [A survey on LLM-as-a-judge](#). *The Innovation*, 7(6):101253.
- Xin Guo, Haotian Xia, Zhaowei Liu, Hanyang Cao, Zhi Yang, Zhiqiang Liu, Sizhe Wang, Jinyi Niu, Chuqi Wang, Yanhui Wang, Xiaolong Liang, Xiaoming Huang, Bing Zhu, Zhongyu Wei, Yun Chen, Weining Shen, and Liwen Zhang. 2025. [FinEval: A Chinese financial domain knowledge evaluation benchmark for large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6258–6292, Albuquerque, New Mexico. Association for Computational Linguistics.
- Xiaohui Hou and Rui Yang. 2021. [Policy signaling and stock price synchronicity: Evidence from china](#). *Journal of International Financial Markets, Institutions and Money*, 75:101355.
- Peter Jansen, Kelly J. Smith, Dan Moreno, and Huitzilin Ortiz. 2021. [On the challenges of evaluating compositional explanations in multi-hop inference: Relevance, completeness, and expert ratings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7529–7542, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.
- Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. [Benchmarking cognitive biases in large language models as evaluators](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545, Bangkok, Thailand. Association for Computational Linguistics.

- Yurina Kosai, Yucheng Xie, Rikuto Tsuchida, and Takehito Utsuro. 2026. [Llm-as-a-judge evaluation of financial news articles generated based on factors of stock price fluctuation](#). In *The 7th Financial Narrative Processing Workshop*, pages 106–113, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Mahnaz Koupaee, Xueying Bai, Mudan Chen, Greg Durrett, Nathanael Chambers, and Niranjan Balasubramanian. 2025. [Causal graph based event reasoning using semantic relation experts](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26169–26199, Vienna, Austria. Association for Computational Linguistics.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhat-tacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. [From generation to judgment: Opportunities and challenges of LLM-as-a-judge](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.
- 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, Singapore. Association for Computational Linguistics.
- Ying Nie, Binwei Yan, Tianyu Guo, Hao Liu, Haoyu Wang, Wei He, Binfan Zheng, Weihao Wang, Qiang Li, Weijian Sun, Yunhe Wang, and Dacheng Tao. 2025. [CFinBench: A comprehensive Chinese financial benchmark for large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 876–891, Albuquerque, New Mexico. Association for Computational Linguistics.
- Shunsuke Nishida and Takehito Utsuro. 2025. [Generating financial news articles from factors of stock price rise / decline by LLMs](#). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 184–195, Abu Dhabi, UAE. Association for Computational Linguistics.
- Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. [Judging the judges: A systematic study of position bias in LLM-as-a-judge](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Guijin Son, Haneul Jung, Moonjeong Hahm, Keonju Na, and Sol Jin. 2023. [Beyond classification: Financial reasoning in state-of-the-art language models](#). In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 34–44, Macao. -.
- Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. [Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges](#). In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 404–430, Vienna, Austria and virtual meeting. Association for Computational Linguistics.
- Meiyun Wang, Kiyoshi Izumi, and Hiroki Sakaji. 2024. [LLMfactor: Extracting profitable factors through prompts for explainable stock movement prediction](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3120–3131, Bangkok, Thailand. Association for Computational Linguistics.
- Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. [Self-preference bias in LLM-as-a-judge](#). In *Neurips Safe Generative AI Workshop 2024*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems*, pages 46595–46623.

A Evaluation Criteria

This section provides the detailed scoring criteria used in the proposed evaluation framework, as summarized in Table 9.

Table 9: Detailed scoring criteria for the proposed evaluation framework.

Dimension	Scoring Criteria
Relevance	5: Almost all content contributes to explaining the stock price movement. The explanation is focused, with no significant irrelevant information, excessive background description, or key unsupported content. 4: Most content contributes to explaining the stock price movement, with only a small amount of weakly related information or minor unsupported content. 3: The article contains both explanatory content and substantial background information or event repetition. Some content has limited explanatory value, or part of the explanation relies on information insufficiently supported by the input. 2: Only a small amount of content actually explains the stock price movement. Most of the article consists of repeated input information, company background, or content with limited explanatory value, or contains substantial unsupported content. 1: The article provides almost no effective explanation of the stock price movement, or the main explanation relies heavily on information unsupported by the input.
Completeness	5: The article provides a complete explanation of how the event affected the company and ultimately led to the stock price movement, with no significant missing causal links. 4: The article provides a basically complete causal chain, with only minor omissions. 3: The article explains part of the causal process but contains noticeable gaps and is not sufficiently complete. 2: The article describes only isolated causes and omits multiple key causal links, preventing a complete explanation. 1: The article provides almost no effective causal explanation.
Economic Plausibility	1 (Plausible): The causal explanation is consistent with general economic knowledge and financial-market reasoning and can be accepted as a reasonable market explanation. 0 (Implausible): The causal explanation clearly violates common economic knowledge or financial-market reasoning, contains a clear logical error or forced causal connection, or relies on a key premise unsupported by the input information. NE (Not Evaluable): The explanation is too brief, lacks key causal links, or requires specialized knowledge to determine its plausibility, making it impossible to reliably evaluate based on general economic knowledge alone.

B Article Generation Prompts

We used two generation settings for producing the financial news articles: Direct and Chain-of-Thought (CoT). In the Direct setting, the article was generated directly from the provided input information. In the CoT setting, the generation was performed in two stages: the model first generated the causal reasoning and then used the reasoning to produce the final news article. All prompts below are translated into English from the original Chinese prompts used in the experiments.

B.1 Direct Prompt

The following prompt was used to generate articles under the Direct setting.

You are a professional financial journalist.
Based on the following information, write a news article explaining the reason for the company's stock price movement.

Input Information
Date: {Date}
Company: {Company}
Company Profile: {Company Information}
Stock Prices over the Past Five Days: {Stock Prices}
Event: {Event}

Requirements
1. Describe the recent movement of the company's stock price.
2. Naturally explain how the event affected the company's stock price.
3. The article should include the company name, stock price movement, and analysis of the cause.

4. Describe the situation based only on the provided information. Do not add facts or speculation that cannot be supported by the input information.
 5. Use complete sentences and do not use bullet points.
 6. Use standard written language.
 7. Keep the article at approximately 250 Chinese characters (about 200–300 characters).
- Output only the body of the news article. Do not output a title or any other explanation.

B.2 CoT Prompt

The following two-stage prompts were used to generate articles under the Chain-of-Thought (CoT) setting, including causal reasoning (Stage 1) and article generation (Stage 2).

B.2.1 Stage 1: Causal Reasoning

The following prompt was used in the first stage of the CoT setting to generate a causal reasoning chain.

You are a senior financial analyst familiar with the Chinese A-share market.
Based on the given information, analyze the causal process underlying the stock price movement and generate a plausible causal reasoning chain.

Input Information
Date: {Date}
Company: {Company}
Company Profile: {Company Information}
Stock Prices over the Past Five Days: {Stock Prices}
Event: {Event}

Task
Starting from the given event, gradually derive how the event ultimately led to the company's stock price increasing or decreasing.

Requirements

1. Each step must have a clear causal relationship with the previous step.
2. General financial knowledge may be used for reasoning, but no specific facts unsupported by the input information may be added.
3. Do not fabricate the company's products, business operations, partnerships, research achievements, or other information not provided in the input.
4. The intermediate process may include factors such as industry effects, market sentiment, investor expectations, capital flows, and earnings expectations, but it is not necessary to include all of them.
5. Keep the reasoning chain concise and retain only the key steps necessary to explain the stock price movement.
6. Limit the reasoning chain to 3–6 steps.

Output Format
Step 1:
Step 2:
...
Final Result: The company's stock price increased (or decreased).

B.2.2 Stage 2: Article Generation

The following prompt was used in the second stage of the CoT setting to generate the final news article from the input information and the causal reasoning chain.

You are a professional financial journalist.
Based on the following input information and causal reasoning chain, write a news article explaining the reason for the company's stock price movement.

Input Information
Date: {Date}
Company: {Company}
Company Profile: {Company Information}
Stock Prices over the Past Five Days: {Recent Stock Prices}
Event: {Event}
Causal Reasoning Chain: {Causal Reasoning Chain}

Requirements

1. Describe the recent movement of the company's stock price.
2. Naturally explain how the event affected the company's stock price.
3. The article should include the company name, stock price movement, and analysis of the cause.
4. Describe the situation based only on the input information and the given causal reasoning chain. Do not add new facts or speculation.
5. Use complete sentences and do not use bullet points.

6. Use standard written language.
7. Keep the article at approximately 250 Chinese characters (about 200–300 characters).
Output only the body of the news article. Do not output a title or any other explanation.

C Evaluation Prompts

We evaluated the generated articles using two prompting settings: Zero-shot and Few-shot. In the Zero-shot setting, the evaluation model was provided with the evaluation instructions, input information, generated article, and evaluation criteria. In the Few-shot setting, the same evaluation prompt was supplemented with manually selected examples illustrating different evaluation scores and reasoning patterns. The prompts below were originally written in Chinese and are presented here in English translation for readability.

C.1 Zero-shot Prompt

The following prompt was used for the Zero-shot evaluation.

You are a senior financial analyst familiar with the Chinese A-share market and experienced in stock research. Your task is to evaluate the quality of the explanation for stock price fluctuations in a financial news article generated by a large language model.

Please strictly follow the requirements below:

1. Evaluate the article only based on the provided input information. Do not use external information about the specific events, companies, or stock price movements being evaluated.
2. Evaluate only the explanation of the reasons for the stock price movement. Do not evaluate the language, writing style, news structure, or wording of the article.
3. The article is required to generate explanations only from the provided information and must not introduce specific facts unsupported by the input. Common financial-market transmission mechanisms may be regarded as reasonable inferences rather than newly introduced facts.
4. General financial-market transmission mechanisms that can be derived from the input information, such as industry conditions, market sentiment, investor expectations, and sector-wide effects, may be regarded as reasonable inferences rather than newly introduced facts.
5. The same stock price movement may have multiple plausible explanation paths. Any explanation that can be reasonably derived from the input information should be evaluated accordingly.
6. Complete the analysis internally and output only the final result. Do not output the analysis process.
7. Output only valid JSON. Do not output any other text, Markdown, explanation, comments, or code blocks.

Input Information

Date: {Date}

Company: {Company}

Company Profile: {Company Information}

Stock Prices over the Past Five Days: {Stock Prices}

Event: {Event}

Generated News: {Generated News}

Evaluation Criteria

Relevance (1–5)

Evaluate whether the content used to explain the stock price movement actually serves an explanatory role.

Focus on whether the article explains the recent stock price movement based on the event and company information, rather than simply repeating the event, listing company background information, or including information with little explanatory value.

Also consider whether the generated explanation remains faithful to the input information. If an explanation relies on specific company facts, such as business activities, products, or production capacity, that are not provided in the input and cannot be reasonably derived from it, the Relevance score should be reduced even if such information would otherwise be relevant to explaining the stock price.

This criterion evaluates whether the explanation is effective and consistent with the input information. It does not evaluate whether the explanation is economically reasonable.

5: Almost all content contributes to explaining the stock price movement. The explanation is focused, with no significant irrelevant information, excessive background description, or key unsupported content.

4: Most content contributes to explaining the stock price movement, with only a small amount of weakly related information or minor unsupported content.

3: The article contains both explanatory content and substantial background information or event repetition. Some content has limited explanatory value, or part of the explanation relies on information insufficiently supported by the input.

2: Only a small amount of content actually explains the stock price movement. Most of the article consists of repeated input information, company background, or content with limited explanatory value, or contains substantial unsupported content.

1: The article provides almost no effective explanation of the stock price movement, or the main explanation relies heavily on information unsupported by the input.

Completeness (1–5)

Evaluate whether the article sufficiently explains how the event affected the company and ultimately led to the recent stock price movement.

Focus on whether the causal chain is complete rather than on the number of causal steps. When an event involves clear intermediate transmission mechanisms, omitting key causal links should result in a lower score.

This criterion evaluates whether the explanation is sufficiently complete. It does not evaluate whether the reasoning is economically reasonable.

5: The article provides a complete explanation of how the event affected the company and ultimately led to the stock price movement, with no significant missing causal links.

4: The article provides a basically complete causal chain, with only minor omissions.

3: The article explains part of the causal process but contains noticeable gaps and is not sufficiently complete.

2: The article describes only isolated causes and omits multiple key causal links, preventing a complete explanation.

1: The article provides almost no effective causal explanation.

Economic Plausibility (1, 0, NE)

Evaluate whether the causal explanation is consistent with general economic knowledge and basic financial-market reasoning.

The evaluator must not consult external sources and should make the judgment based only on general economic knowledge.

The explanation does not need to have been empirically verified.

This criterion evaluates only whether the explanation is economically reasonable. It does not evaluate whether the explanation is complete.

The evaluation uses three labels:

Plausible (1): The causal explanation is consistent with general economic knowledge and financial-market reasoning and can be accepted as a reasonable market explanation.

Implausible (0): The causal explanation clearly violates common economic knowledge or financial-market reasoning, contains a clear logical error or forced causal connection, or relies on a key premise unsupported by the input information.

Not Evaluable (NE): The explanation is too brief, lacks key causal links, or requires specialized knowledge to determine its plausibility, making it impossible to reliably evaluate based on general economic knowledge alone.

The three evaluation criteria are independent and should be evaluated separately.

- Relevance evaluates whether the article actually explains the stock price movement and whether the explanation is based on information supported by the input. - Completeness evaluates whether the causal explanation is sufficiently complete. -

Economic Plausibility evaluates whether the causal explanation is consistent with general economic reasoning. A score for one criterion should not directly affect the others. For example, an explanation may be highly relevant but economically implausible, or economically plausible but incomplete.

Output Format

Output only the following JSON. Do not add, remove, or modify any fields:

```
{
  "relevance": {
    "score": <1-5>,
    "reason": ""
  },
  "completeness": {
    "score": <1-5>,
    "reason": ""
  },
  "plausibility": {
    "score": "1" | "0" | "NE",
    "reason": ""
  }
}
```

Output Requirements

1. relevance.score and completeness.score must be integers and must be one of 1, 2, 3, 4, or 5.

2. plausibility.score must be "1", "0", or "NE".

3. All reason fields must be written in Simplified Chinese.

4. The length of the reasons is unrestricted, but they should be specific and clear without unnecessary repetition.

5. relevance.reason must identify which parts of the article effectively explain the stock price movement, which parts are background information, event repetition, or information with limited explanatory value, and which parts are insufficiently supported by the input, and explain how these factors affect the Relevance score.

6. completeness.reason must describe the main causal chain formed by the article and clearly identify any missing key causal links.

7. plausibility.reason must explain why the causal explanation is consistent or inconsistent with general economic reasoning, or why its plausibility cannot be reliably determined.

8. The reasons must evaluate the specific content of the generated article and must not simply repeat the scoring criteria.

9. The output must be valid JSON.

10. Do not output anything other than the JSON.

C.2 Few-shot Prompt

The Few-shot evaluation used the same prompt and evaluation criteria as the Zero-shot setting, with manually selected evaluation examples added to the prompt. Each example included the input information, generated article, human evaluation scores, and corresponding evaluation reasons. The examples were selected to cover different score levels and explanation characteristics.

D Agreement Results by Generation Condition

This appendix reports the agreement between each LLM judge and the human evaluation separately for the four model–prompting combinations: Gemini with Direct prompting, Gemini with CoT prompting, GPT with Direct prompting, and GPT with CoT prompting. The four subsets contain 99, 52, 26, and 23 articles, respectively, for a total of 200 articles. For Relevance and Completeness, Pearson’s correlation (r) and Spearman’s correlation (ρ) are reported. For Economic Plausibility (EP), Cohen’s κ is reported.

Table 10: Agreement with human evaluation by generation condition for Claude Opus 5 as the judge.

Model	Prompting	n	Rel (r)	Rel (ρ)	Comp (r)	Comp (ρ)	EP (κ)
Gemini	Direct	99	0.559	0.543	0.441	0.410	0.221
Gemini	CoT	52	0.411	0.361	0.062	0.074	0.568
GPT	Direct	26	0.532	0.544	0.333	0.308	0.366
GPT	CoT	23	0.362	0.324	0.303	0.303	n/a

Table 11: Agreement with human evaluation by generation condition for GPT-5.5 as the judge.

Model	Prompting	n	Rel (r)	Rel (ρ)	Comp (r)	Comp (ρ)	EP (κ)
Gemini	Direct	99	0.517	0.475	0.395	0.363	0.663
Gemini	CoT	52	0.413	0.325	-0.033	-0.041	0.306
GPT	Direct	26	0.319	0.302	0.334	0.281	0.574
GPT	CoT	23	0.586	0.586	0.253	0.253	n/a

Table 12: Agreement with human evaluation by generation condition for Gemini-3.5-Flash as the judge.

Model	Prompting	n	Rel (r)	Rel (ρ)	Comp (r)	Comp (ρ)	EP (κ)
Gemini	Direct	99	0.403	0.393	0.413	0.385	0.499
Gemini	CoT	52	0.471	0.349	n/a	n/a	0.472
GPT	Direct	26	0.510	0.512	0.348	0.368	0.480
GPT	CoT	23	n/a	n/a	n/a	n/a	n/a

Notes on undefined statistics. Some statistics are reported as n/a because the corresponding variable had no variation within the subset. For Gemini-generated articles evaluated by Gemini under CoT prompting, all Completeness scores assigned by the judge were 5, making the correlation with human scores undefined. For GPT-generated articles evaluated by Gemini under CoT prompting, all Relevance and Completeness scores assigned by the judge were 5, so the corresponding correlations were undefined. For the GPT–CoT subset, human EP labels and all judge EP labels were 1, resulting in no class variation and therefore an undefined Cohen’s κ .

The relatively low correlations in some CoT subsets may partly reflect limited variation in both human and judge scores, as many articles received the highest Relevance and Completeness scores from both.

E Confusion Matrices

The complete score confusion matrices between human and LLM evaluations for the three evaluation dimensions are shown in Figure 4.

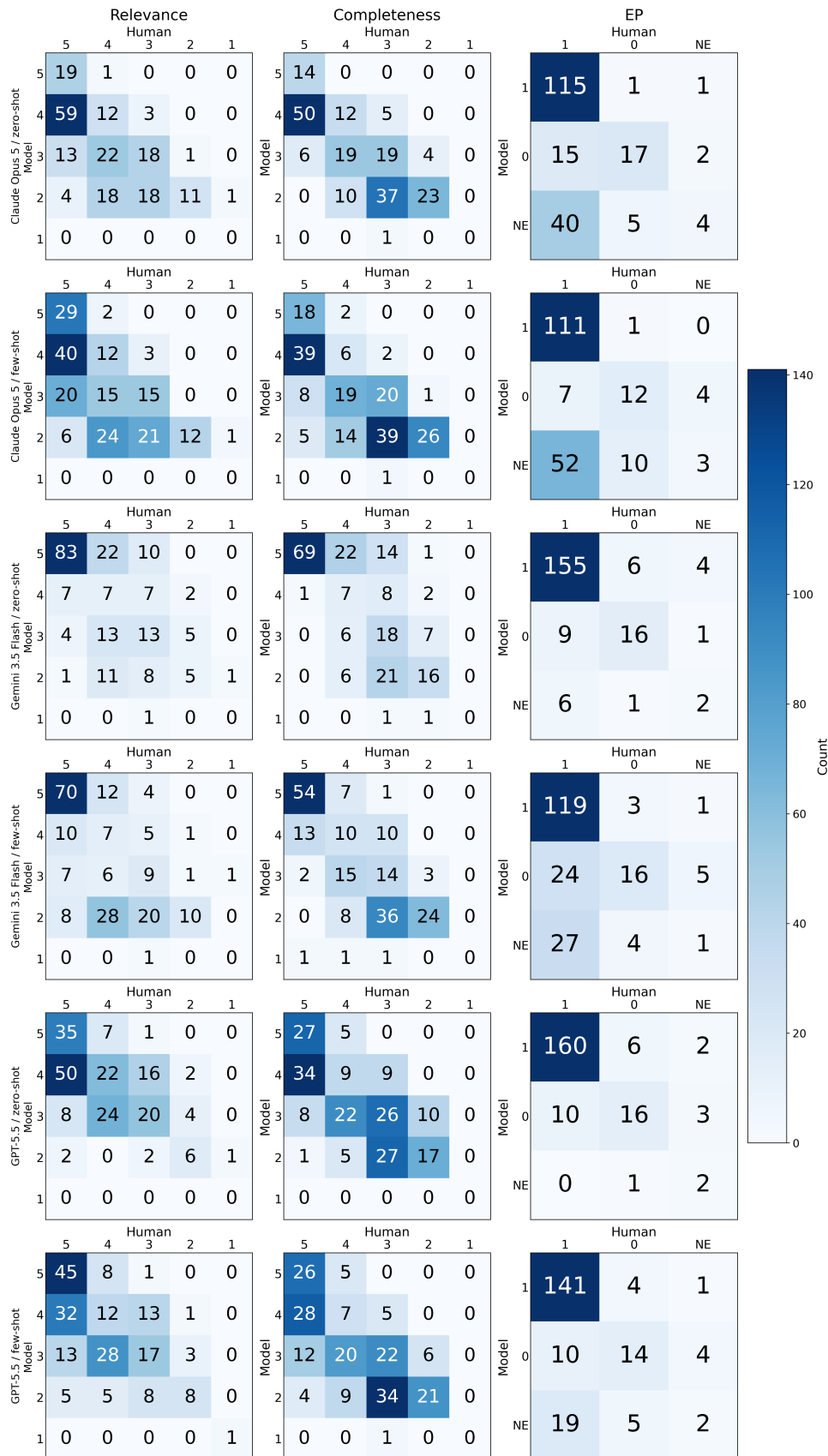


Figure 4: Score confusion matrices between human and LLM evaluations for Relevance, Completeness, and Economic Plausibility.