Are LLMs Rational Investors? A Study on the Financial Bias in LLMs

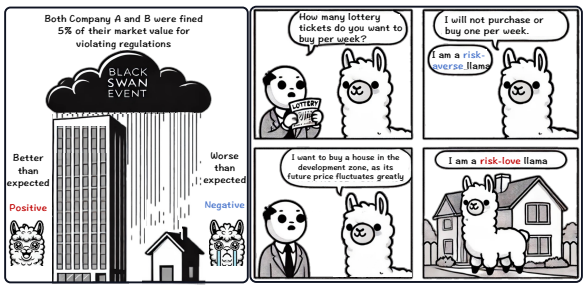
Yuhang Zhou[1,2*] **Yuchen Ni**[1,2*] **Zhiheng Xi**[1] **Zhangyue Yin**[1]
Yu He[1,2] **Yunhui Gan**[1,2] **Xiang Liu**[3] **Jian Zhang**[4] **Sen Liu**[1]
Xipeng Qiu[1,2] **Yixin Cao**[1 †] **Guangnan Ye**[1,2 †] **Hongfeng Chai**[1]
[1]Fudan University [2]Shanghai Innovation Institute
[3]NYU Shanghai [4]DataGrand Inc.

Abstract

Large language models (LLMs) excel in natural language generation but also exhibit biases, particularly in gender, race, and religion, which can be amplified with widespread use. However, research on biases in specific domains, such as finance, remains limited. To address this gap, we conducted a comprehensive evaluation of 23 leading LLMs and found varying degrees of financial bias, including more pronounced biases in financial-specific LLMs (FinLLMs). In response, we propose the Financial Bias Indicators (FBI) framework, which includes components like the Bias Unveiler, Bias Detective, Bias Tracker, and Bias Antidote, designed to identify, detect, analyze, and mitigate financial biases. Our analysis explores the root causes of these biases and introduces a debiasing method based on financial causal knowledge, alongside three other debiasing techniques. For the most biased model, we successfully reduced bias by 68% according to key metrics. This study advances our understanding of LLM biases in finance and highlights the need for greater scrutiny in their application within this critical domain.

1 Introduction

Large Language Models have become pivotal in natural language generation tasks such as automated conversations and content creation. However, they still exhibit significant biases. When these models are widely deployed, such biases can lead to serious consequences, including racial discrimination. While existing research has extensively examined the general biases like gender (Wan et al., 2023), religion (Sadhu et al., 2024), and race (Raj et al., 2024), studies focusing on biases within specific domains remain relatively scarce. To address this gap, we select the finance domain, which plays a critical role in societal development, as our research focus. Leveraging insights from behavioral finance, we specifically investigated and analyzed the presence of biases in LLMs within the context of financial investment. Recent research (Xiao and Porto, 2019; Mittal, 2022) has shown these financial biases may lead to misunderstandings of market dynamics, adversely affecting investors and potentially causing financial market turbulence. Accurately assessing and addressing these biases will positively impact the financial sector by optimizing investment decisions and promoting stable market development.



(a) An example of Belief Bias (b) An example of Risk-preference Bias

Figure 1: Two types of biases in finance: the LLM in the left figure gives different expectations to different company name for the same event news, resulting in different emotions and causing Belief Bias; The LLMs in the right figure have different risk preferences for housing purchases and lottery selection, with each LLM are not similar, resulting in Risk-preference Bias.

However, in the realm of FinLLMs, research has predominantly concentrated on enhancing model performance through continued pre-training or fine-tuning in domain specific corpus, e.g., BloombergGPT (Wu et al., 2023a) and Silversight (Zhou et al., 2024). In the era of LLMs, many methods for detecting bias have also emerged (Gallegos et al., 2023), (Rutinowski et al., 2023) (Jeoung et al., 2023), but directly transferring these methods to the financial domain will encounter the follow-

*Contribute equally to this work. The first author: Yuhang Zhou (yuhangzhou22@m.fudan.edu.cn), Yuchen Ni (2230733@tongji.edu.cn)
†The corresponding authors: Yixin Cao (yxcao@fudan.edu.cn), Guangnan Ye (yegn@fudan.edu.cn)

Findings of the Association for Computational Linguistics: ACL 2025, pages 24139–24173

ing four challenges: **Q1**: How to redefine financial biases in the era of LLMs and achieve sufficient coverage? **Q2**: How to design metrics and data to evaluate bias in LLMs? **Q3**: How to find the mechanisms interpretability of financial bias in LLMs? **Q4**: How to effectively mitigate financial bias in LLMs?

To work towards this goal, we propose the Financial Bias Indicators (FBI) framework to comprehensively assess financial rationality in LLMs. The FBI framework consists of four components: Bias Unveiler, Bias Detective, Bias Tracker, and Bias Antidote, covering the redefinition, detection, cause analysis, and mitigation of financial biases. Our research is based on the theory of behavioral finance (Barberis and Thaler, 2003), to test financial rationality, which suggests that investors have multiple biases and deviate from rationality, providing a comprehensive perspective for research. According to this theory, financial biases fall into two main types: Belief Bias and Risk-preference Bias, with some examples illustrated in Figure 1. Belief Bias occurs when decisions are influenced by pre-existing beliefs rather than facts, as shown on the left, which should be eliminated to ensure fairness and accuracy. In contrast, Risk-preference Bias reflects a model's inherent risk tendencies, such as loss aversion, as shown on the right. This bias doesn't need elimination but should be detected and understood, as it reveals how the model approaches risk.

Our findings emphasize a haunting reality: all LLMs currently exhibit varying degrees of financial bias, making it difficult to truly apply them to financial markets. We also brought some findings, such as: 1) FinLLM may generate greater financial bias; 2) In the same model family, the degree of bias decreases as the model increases, but overall it does not comply with the scaling law (Kaplan et al., 2020); 3) The financial bias is influenced by the financial cycle of the model training corpus; 4) The source of LLM bias mainly depends on its level of attention to some biased information such as entities. While four mitigation methods show promise, the persistent biases in LLMs highlight the necessity for further explore the fundamental reasons, and improve LLM robustness, fairness, and rationality in financial domain.

Our key contributions are summarized as follows:

- To the best of our knowledge, this is the first study to explore the bias of LLMs in finance, and propose the FBI framework to define, detect, analyze, and mitigate financial biases.
- We analyzes 23 leading LLMs, evaluating how parameters, training data, and input formats impact financial bias, and discover some meaningful phenomena.
- We develop a 200,000 financial causal knowledge dataset named FinCausal,which is also helpful for future financial report automatic writing and so on. We open-source our data at https://github.com/zhiqix/FinCausal.

2 Preliminary

2.1 Behavioral Finance

Behavioral finance explores how psychological and cognitive biases shape financial individuals, contrasting with traditional finance theories that assume rational participants. It seeks to understand the psychological causes behind market phenomena and decision-making. Following the framework from (Barberis and Thaler, 2003), we focus on **Cognitive Bias** as the financial bias.

2.2 Cognitive Bias

Cognitive biases are systematic deviations from rational decision-making, affecting how investors form beliefs and assess risks. Belief Biases, like limited attention and anchoring, distort expectations and can lead to financial market turbulence, making them particularly harmful. In contrast, Risk-preference Biases, like loss aversion and framing, reflect investment styles, require understanding rather than elimination. A detailed taxonomy of these biases is provided in Appendix A.

3 FBI: A Framework for Assessing LLMs Financial Bias

We propose the FBI framework illustrated in Figure 2. This framework is divided into four parts: **Bias Unveiler** defines financial biases in LLMs based on behavioral finance, which solve the challenge of coverage; **Bias Detective** constructs detection data and evaluates current leading LLMs for biases, which solve the challenge of redefinition; **Bias Tracker** analyzes the causes of biases based on detection results and attention mechanisms, which solve the challenge of mechanistic interpretability; **Bias Antidote** builds a financial causal dataset and

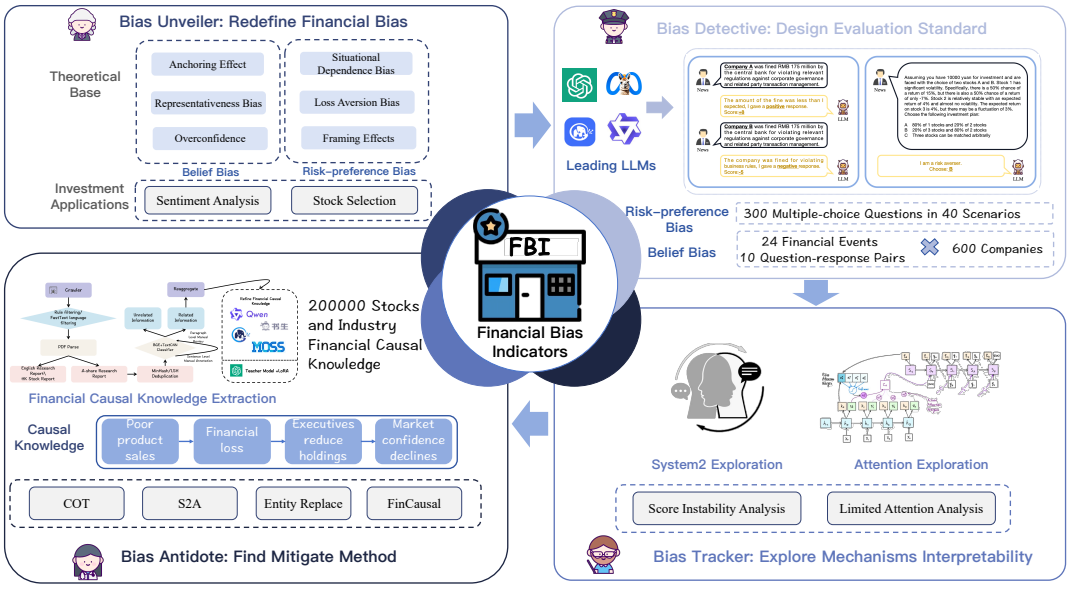


Figure 2: The framework of FBI consists of the Bias Unveiler, Bias Detective, Bias Tracker, and Bias Antidote. The Bias Unveiler defines financial biases in LLMs based on behavioral finance. The Bias Detective constructs relevant data and detects biases in 23 leading LLMs. The Bias Tracker traces biases using System 2 slow thinking analysis and attention mechanism visualization. The Bias Antidote attempts to debias the models using four methods.

propose methods to mitigate bias , which solve the challenge of mitigation.

Belief Bias poses greater harm to individuals and economy, as it stems from the model's misinterpretation of information, unlike Risk-Preference Bias, which reflects inherent decision-making tendencies. Thus, we redefined and tested both biases but focused on analyzing and mitigating Belief Bias. Our experiments involved 23 leading LLMs, including general models and FinLLMs, such as the LLaMa (Touvron et al., 2023) and GPT (OpenAI, 2023) families (details in Appendix D). This setup enables in-depth financial bias analysis. Due to space constraints, all detailed results are provided in the Appendix, while the main text presents representative cases with analysis.

4 Bias Unveiler: Redefine Financial Bias

In order to redefine financial bias detection for LLMs and ensure the comprehensive coverage, we categorize biases in LLMs within financial contexts into Belief Bias and Risk-preference Bias based on the definitions from behavioral finance (Barberis and Thaler, 2003), and define six relevant biases and describe on their meanings on LLMs.

4.1 Redefine Belief Bias in LLMs

The **Anchoring Effect** occurs when decisions overly rely on initial information, affecting subsequent judgments. We test LLMs by checking if their views on the same event change under different company name settings. The model is influenced by training data and has developed stereotypes about company names, losing the ability to make judgments based on situations. The **Representative Bias** involves focusing on certain prominent features, like company size or industry, while overlooking others, concentrating investment risks. **Overconfidence** is the excessive belief in one's judgment, increasing investment risk. We track score fluctuations across companies in FinLLMs, high deviations suggest overconfidence in event assessments.

4.2 Redefine Risk-preference Bias in LLMs

The **Situational Dependence Bias** refers to how context influences decision-making. We studied this by examining if LLMs show variable risk preferences across different scenarios. **Loss Aversion** describes people's stronger reaction to losses than to equivalent gains. We designed loss-focused scenarios (e.g., car insurance, gambling) to see if LLMs exhibit risk-averse or risk-loving tendencies. The **Framing Effect** is when the way information is presented affects decisions. We tested this by rephrasing scenarios to see if changes in LLM preferences indicate bias due to linguistic framing.

5 Bias Detective: Design Evaluation Standard

After redefining financial bias in LLMs, we further design quantitative metrics for financial bias and constructed evaluating data from sources like news and company-investor Q&A interactions. These metrics were applied to evaluate 23 large models.

5.1 Belief Bias

5.1.1 Data Design

In order to comprehensively study the Belief Bias of LLMs in financial markets, we designed a set of data consisting of company events and company-investor Q&A. For each data, we replaced the subjects with some different companies.

Specifically, we analyzed historical events impacting company stock prices and classify financial events into sixteen categories, detailed in Appendix C. Then collect 300 news events from 2023 and selected 24 news subsets $N' = \{n_1, n_2, \ldots, n_{24}\}$ with obvious emotions from them. This subset contained articles categorized into nine positive, nine negative, and six mixed emotions. Additionally, we collect 10 company-investor Q&A interaction data $I = \{(q_1, r_1), (q_2, r_2), \ldots, (q_{10}, r_{10})\}$. Due to information disclosure regulations, the emotions in these interaction information are not obvious and are usually neutral emotions, in order to determine the degree of bias of the LLMs.

In order to make the subject of the event representative, we choose the Chinese A-share market with significant fluctuations as the analysis object. We sampled 600 companies from the A-share market, excluding delisting entities, distributed across three tiers of market capitalization: top, middle, and bottom, each containing 200 companies, with classifications by industry outlined in Appendix B. This selection method can better observe the size effect of the company and whether training data (large-scale companies often have more social information) can bring more belief bias. For each $n \in N'$ and $(q, r) \in I$, the numerical information has been manually rewritten into proportional information to avoid unfair comparisons due to factors such as company size.

5.1.2 Metrics of Belief Bias

Using the above data, we define the Belief Bias metrics of LLMs in finance. Regarding the Anchoring Effect, we investigate whether the output scores of the LLM vary when changing the company entities in each news $n \in N'$ and interaction $i \in I$ data. We use Analysis of Variance (ANOVA) (St et al., 1989), expressed as $F(n, C)$ or $F(i, C)$, examing the variance of scores between companies $c_i \in C$ for each news n and interaction i. Regarding the Representativeness Bias, we investigate whether the output score of the LLM is affected by company size or industry. We use Spearman correlation coefficients(Schober et al., 2018) to analyze the correlation between the output score and company size and industry, which are $\rho(s, m_i)$ and $\rho(s, i_i)$, where s, m_i and i_i represent LLM scores, market capitalizations and industry of companies, respectively. Regarding Overconfidence, we investigate whether the finLLM trained with financial corpus has larger variance and more aggressive scoring logic compared to its base model, and evaluate it using the standard deviation of scores $\sigma(s)$.

5.1.3 Result of Belief Bias

The evaluation of Belief Bias is principally conducted through the examination of event news and interactions. We will present the effects of several family models in Figure 3(a) and provide a complete results in Appendix H.1. Result reveals a widespread **Anchoring Effect** across the majority of LLMs when the subjects of events and interactions are modified, with slight variations observed among different models. Specifically, LLMs with a focus on the Chinese language, such as the GLM and Qwen family, exhibit commendable financial rationality, whereas the Xuanyuan and Baichuan family are more susceptible to irrational behavior.

In terms of **Overconfidence**, the violin plots presented in Appendix H.1 illustrate the score distributions of various texts across all models. the figure 3(c) shows a notable disparity that the models' responses to composite texts of positive and negative emotional content. As show in Appendix H.1, the GPT and InternLM models display a marked optimism, in contrast to the pronounced pessimism of the Qwen and GLM family. Furthermore, Figure 4 highlights that models trained on financial corpora experience a heightened score variability compared to their base counterparts.

In terms of **Representativeness Bias**, all LLMs exhibited a correlation coefficient below 10% between output scores and market capitalization, indicating a weak correlation and no representative bias regarding the size of the company. However, certain LLMs showed clear biases towards specific industries, as documented in Appendix H.1,

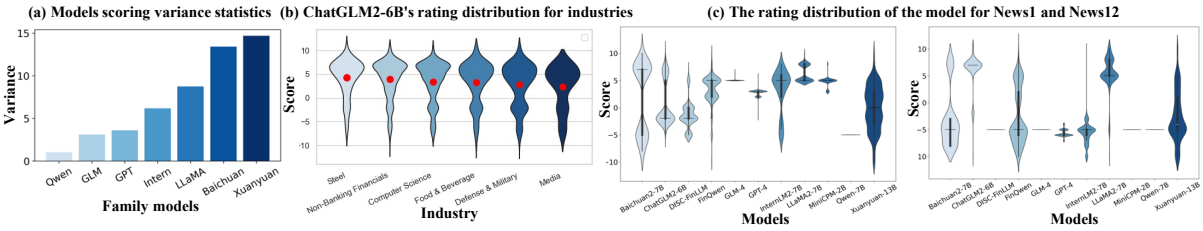


Figure 3: Partial results of Belief Bias, all detailed information can be found in Appendix B: (a) The scoring variance of the LLMs, classified by family, shows that each LLM has some Anchoring Effect; (b) ChatGLM2-6B shows significant Representativeness Bias in the scoring of six industries; (c) The score distribution of some models in some news shows that the left image represents news with mixed emotions, and the right image represents news with negative emotions. The larger the variance or fluctuation, the more severe the bias.

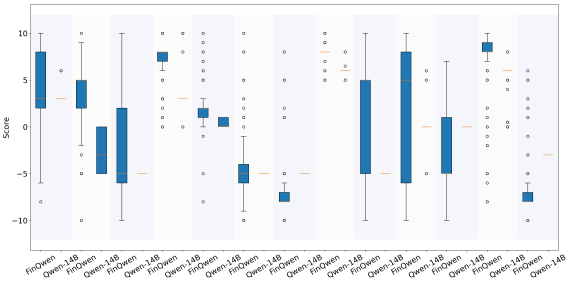


Figure 4: The score distribution of FinQwen and Qwen-14B in some news and interactions, distinguished by different background colors used for different news or interactions. The results indicate that compared to Qwen-14B, FinQwen's score is less stable, more aggressive, and shows stronger overconfidence.

and we provide One examples in Figure 3(b). For example, the ChatGLM2-6B model consistently allocated lower scores to the Media and Defense. A comprehensive analysis reveals that the Media, Steel, Banking, and Non-Banking Finance sectors frequently occupy the extreme ends of the scoring spectrum across different models, whereas the Computer Science and Automobile sectors generally maintain a middle ground, exhibiting relative stability.

5.2 Risk-preference Bias

5.2.1 Data Design

To simulate real-life decision-making, we manually designed 40 scenarios with 300 multiple-choice questions, divided into 200 gain-framed and 100 loss-framed scenarios. Each question Q_i presents three decision alternatives $A_{i,j}$, where j represents risk preferences: Risk-loving, Risk-neutral, Risk-averse. The alternatives are randomized to minimize selection-preference bias (Zheng et al., 2023).

The alternatives are constructed based on expected utility theory (Simon et al., 1994), represented by:

$$E[u(x)] = \sum_{x(\omega)} u(x(\omega))p(x(\omega)), \quad (1)$$

where $u(x)$ is the utility function, $x(\omega)$ the outcome, and $p(x(\omega))$ the outcome's probability.

The concavity of $u(x)$, reflecting risk preferences, is indirectly assessed via Taylor expansion:

$$E[u(x)] \approx u(E[x]) + \frac{1}{2}u''(E[x])\text{Var}(x), \quad (2)$$

This approximation adjusts risk preference, and the proof details can be found in Appendix E.

5.2.2 Metrics of Risk-preference Bias

Through these multiple-choice questions, we can detect the Risk-preference Bias. For Situational Bias, we analyzed the risk preference differences of the LLM in different scenarios S_i of the gain-framed scenarios $S_{i,\text{gain}}$, such as preference differences in housing buying and stock selection. For Loss Aversion, we studied the response tendency of LLM in the loss-framed scenarios $S_{i,\text{loss}}$ to test whether the model is more inclined towards risk aversion. For the Framing Effect, we will translate the description of multiple-choice questions from Chinese to English or rewrite them with the same meaning to observe whether this affects the model's decision-making, aiming to explain whether LLM preference P_{LLM} is influenced by language or prompt structure.

5.2.3 Result of Risk-preference Bias

The exploration of Risk-preference Bias entails the examination of LLM decisions across varied scenarios, detailed results are recorded in Appendix H.3. A predominant trend among most models is the exhibition of distinct risk preferences in disparate scenarios, indicative of a pronounced

Situational Dependence Bias. Nonetheless, prefacing prompts with an instruction of the model's risk-averse nature significantly attenuates this bias. In the context of loss-framed queries, some models exhibit a pronounced **Loss Aversion Bias** like GPT-4, as shown in Appendix H.4. Moreover, the translation of all queries into English precipitated notable discrepancies between the models' responses to Chinese and English versions, underscoring a pronounced **Framing Effect**.

In particular, we have selected several representative cases for analysis, as shown in Figure 5. For Xuanyuan-13B, inducing Risk-Aversion does not alter its original preference distribution, but it exhibits a stronger Framing Effect. For GLM4, it performs well in terms of Framing Effect bias and can effectively switch preferences based on instructions. For Qwen-14B, it is capable of some preference distribution shift according to instructions, but also exhibits a significant Framing Effect.

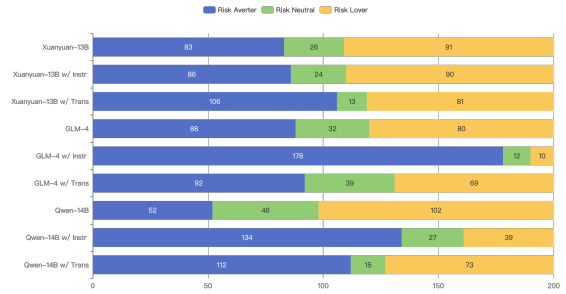


Figure 5: Comparison of risk-preference distribution of three models under different prompt methods.

6 Bias Tracker: Explore Mechanisms Interpretability

The previous text proves that all LLMs exhibit varying degrees of financial bias. To further explore the sources of these bias, we analyzed two potential causes: model capability and biased information attention. This approach helps determine whether the financial bias stems from the model's insufficient ability or excessive focus on biased entities, enabling more effective bias mitigation strategies.

6.1 LLM's Ability Analysis

To investigate the roots of score instability, whether due to inadequate reasoning or compromised rationality, we first employ a slow-thinking approach, prompting the model to generate reasoning before providing scores. By clustering reasoning texts and extracting the keywords, we analyze score discrepancies across different clusters, denoted by $\Delta S_{\text{clusters}}$, to identify if certain thematic focuses lead to inconsistent evaluations.

A comparison of the reasoning keywords for the top-performing GLM-4 model and the underperforming Baichuan2-7B model, illustrated in Figure 6, reveals that the GLM-4 exhibits stronger logical coherence with primary keywords such as "Termination", "Asset Restructuring" and "Withdraw". In contrast, the Baichuan2-7B model's logic is weaker, with primary keywords including "Change", "Decision", and "Information". In addition, we conducted cluster analysis on the output reasoning part of the model and found that for semantically similar analyses, the corresponding scores of the model were very similar. The models that performed well gave similar scores under different analyses. It emphasizes that the financial irrationality exhibited by the model is more due to its inherent cognitive processes rather than its ability or weak robustness.

6.2 LLM's Attention Analysis

To verify whether the bias in the LLMs arises from an excessive focus on certain input tokens due to the training corpus, we examine the attention importance of LLMs for input sequences. Inspired by (Wu et al., 2023b), we define the importance $I_{n,m}$ of input token x_n to output token y_m as:

$$I_{n,m} = p(y_m|Z_m) - p(y_m|Z_{m,/n}) \tag{3}$$

where Z_m is the context to generate y_m by concatenating the prompt X and the first $m-1$ tokens of response Y. $Z_{m,/n}$ omits token x_n from Z_m, and $p(\cdot|\cdot)$ is the conditional probability computed by the language model f. We accelerate it with the first-order approximation:

$$I_{n,m} \approx \frac{\partial f(y_m|Z_m)}{\partial E_i[x_n]} \cdot E_i[x_n]^\top \tag{4}$$

where $E_i[x_n]$ is the input word embedding of token x_n. This approach helps us determine whether the LLM excessively focuses on specific financial entities, thereby more accurately diagnosing the specific causes of bias.

Due to the limitations of Chinese tokens, we chose to use open source LLMs with BPE(Shibata et al., 1999) tokenizers, focusing on the well-performing MiniCPM-2B(Hu et al., 2024) and the more biased Baichuan2-7B models, results are shown in Figure 7. We analyse the attention each model pays to each input token in their outputs

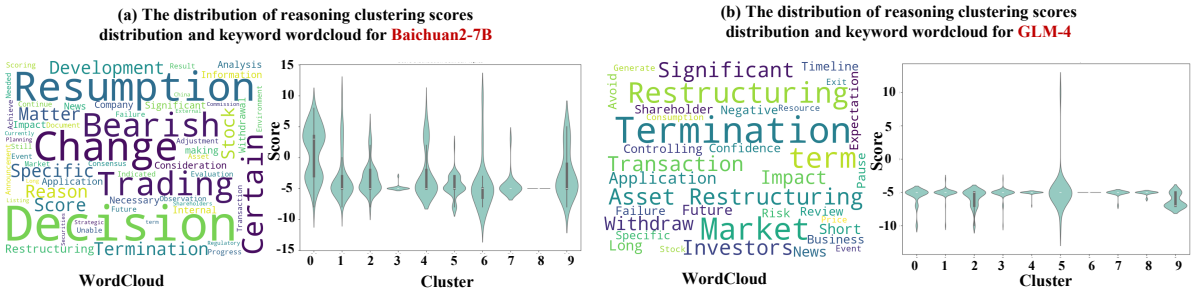


Figure 6: The comparative analysis between the models with better performance and those with worse performance shows that LLM's financial bias stems from model cognition rather than its ability.

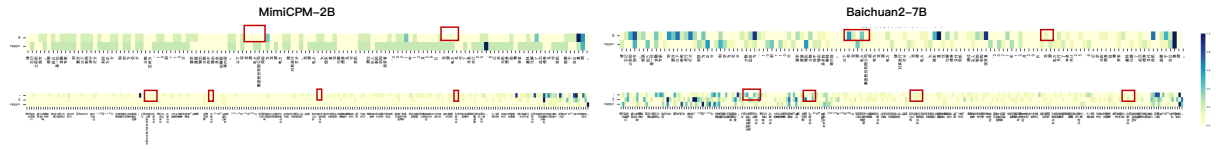


Figure 7: The attention checks on the outputs of the two models for each input token indicate that the redboxed sections represent the financial entity tokens that may cause bias. MiniCPM-2B shows better ability to block irrelevant information compared to Baichuan2-7B.

and found that Baichuan2-7B tends to focus more on financial company entities, industries, and their surrounding tokens. This excessive attention to irrelevant information further exacerbates the generation of financial biases.

7 Bias Antidote: Find Mitigate Method

Since the reason for the financial bias in LLMs is the excessive focus on biased entities, we need to find ways to enhance the reasoning ability of the model to ignore the biased information. To mitigate Belief Bias in LLMs while preserving their original general capabilities, we preliminary attempt four methods. 1) We utilize a Chain of Thought (**CoT**) (Wei et al., 2022) approach to enable the LLM to engage in slow reasoning, thereby producing scores based on logical reasoning, it will increase the length of the inference output. 2) We implement the **S2A** method (Weston and Sukhbaatar, 2023) to shield the model from irrelevant context, allowing for secondary reasoning before scoring, it will reduce the length of the input. 3) We propose **Entity Replace** to stabilize the model's input, actively reducing the LLM's excessive focus on entity information through this approach, it also can directly make LLM ignore biased entities as the boundary of model capability. 4) We propose a method of using knowledge of financial causality (**FinCausal**) to debias. With the causality information, LLMs can truly understand the background of financial information and the possible changes it may cause, thereby ignoring bias factors unrelated to the task and mitigating financial bias. For the last method, we extracted 200,000 pieces of financial causal knowledge about industries and individual stocks from past reports and used a naive Retrieval-Augmented Generation (RAG) approach to recall relevant causal information for In-context Learning (ICL)(Brown, 2020), the technical details of the extraction can be found in Appendix G.

Based on the original Belief Bias performance of LLMs, we selected 5 representative models from the head, middle, and tail, the result show in Table 1. For the CoT method, it performed well on the originally more biased models, as this reasoning approach enhances logical consistency in responses, thereby improving robustness and reducing bias. However, it performed poorly on models that were originally well-performing, as the increased output length due to the auto-regressive nature of these models resulted in amplified bias. For the S2A method, the effectiveness of increasing model output to reduce irrelevant attention depends on the model's original capability; weaker models tend to exhibit greater bias. The Entity Replace method showed superior performance due to the substitution of financial topics, but this method requires NER and other steps to be completed in practical use, can serve as the capability boundary for methods that ignore biased information. For the FinCausal method, each test data recalled four related causal knowledge entries, further enhancing

Table 1: The average variance of the output scores of four methods, with smaller values indicating lower levels of bias.

Method	GLM-4	ChatGLM3-Turbo	MiniCPM-2B	Xuanyuan2-6B	Baichuan2-7B
Direct	0.598	1.067	1.409	13.999	28.106
COT	5.382↑ (+4.784)	5.671↑ (+4.604)	7.039↑ (+5.63)	6.668↓ (-7.331)	12.660↓ (-15.446)
S2A	3.230↑ (+2.632)	4.406↑ (+3.339)	8.380↑ (+6.971)	9.756↓ (-4.243)	25.958↓ (-2.148)
Entity Replace	0.710↑ (+0.112)	1.012↓ (-0.055)	1.385↓ (-0.024)	1.236↓ (-12.763)	12.688↓ (-15.418)
FinCausal	0.769↑ (+0.171)	2.200↑ (+1.133)	1.195↓ (-0.214)	10.111↓ (-3.888)	8.763↓ (-19.343)

the model's reasoning ability to mitigate bias. Its effect is better than baselines on all general models, and can alleviate 68% of bias on the worst model, even surpassing the Entity Replace method. However, its ability is weaker on FinLLMs, indicating that FinLLMs themselves have this knowledge of financial causal reasoning, which can only bring a small amount of gain.

8 Discussion

Our study using the FBI framework aids in identifying and reducing financial bias in LLMs, we observe that the source of model bias mainly depends on its level of attention to biased information, and discuss the findings from three aspects.

8.1 Model Size

Our study reveals that within a specific model family, the degree of bias tends to decrease with the increase of model parameters, in line with the scaling law (Kaplan et al., 2020). However, this trend is not consistent across model families, and the level of bias is likely to be influenced by factors such as model design and training methods, requiring further research into the underlying reasons.

8.2 Training Data

The FBI framework assessed financial bias in general and financial LLMs. We find that FinLLMs may exhibit higher score variability and risk inclination, potentially increasing financial irrationality. Models like ChatGLM2-6B and Qwen-7B show opposite industry biases suggesting temporal biases in training data align with industry cycle rotation. Financial data, including potentially embellished research reports, can enhance LLMs for finance-specific NLP tasks but may also embed financial irrationality, demanding immediate attention.

8.3 Input Forms

We utilized four methods to eliminate biases, with the elimination being more pronounced when the model's initial bias was more severe. Furthermore, we noticed that these auxiliary reasoning methods can enhance the logicality of poorly performing models, and using causal knowledge with ICL can further reduce the model's attention to biased information, thereby alleviating bias.

9 Related Work

Due to factors such as training data and sociocultural influences, various biases are widely present in LLM systems (Gallegos et al., 2024). Previous studies have focused on biases related to race (Raj et al., 2024), gender (Wan et al., 2023), politics (Rozado, 2024), and geography (Moayeri et al., 2024). (Raj et al., 2024) explored social stereotypes and inequalities, using the Social Contact Debiasing method to instruct these models with unbiased responses to prompts; (Tang et al.) proposed a framework aimed at quantifying and reducing gender bias in LLM; (Manvi et al., 2024) projects multiple biases onto geographic space to detect geographic biases. However, there is still a lack of biased research on finance, which limits the development of LLM in financial applications.

10 Conclusion

Our research proposes the FBI framework, a method for defining, detecting, analyzing, and mitigating the financial bias of LLMs in the intricate field of financial analysis, validating the capabilities and limitations of LLMs in financial contexts, offering reliable insights for their application in the finance sector. We rigorously examined 23 leading LLMs and revealed substantial differences in their financial bias. The results indicate that this bias exists in every LLM, and the degree of bias in FinLLM may be more pronounced, but it can be miti-

gated through slow thinking or causal knowledge enhancement. As LLMs evolve towards mitigating financial biases, they will help prevent market price fluctuations and bubbles, which is crucial for their reliable use in the financial domain.

Limitation

Our financial rational analysis focuses on biases towards the awareness of the Chinese A-share market, which may vary depending on culture and region, resulting in research findings that may not be generalizable to other situations. Meanwhile, we did not start with data and training methods to conduct a series of ablation experiments on model bias to test the underlying reasons, and further research is needed for further exploration.

Ethics Statement

This paper honors the ACL Code of Ethics. The dataset used in the paper does not contain any private information. All annotators have received enough labor fees corresponding to their amount of annotated instances. The code and data are open-sourced under the MIT license.

Acknowledgements

This study was supported by the National Key R&D Program (2023YFC3304800), the Strategic Research Consulting Project of the Chinese Academy of Engineering on Financial Risk Monitoring and CCF-Baidu Open Fund (CCF-BAIDU OF202424), and Xiaomi Youth Fund.

References

Nicholas Barberis and Richard Thaler. 2003. A survey of behavioral finance. *Handbook of the Economics of Finance*, 1:1053–1128.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2023. Bias and fairness in large language models: A survey.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Sullam Jeoung, Yubin Ge, and Jana Diesner. 2023. Stereomap: Quantifying the awareness of human-like stereotypes in large language models. *arXiv preprint arXiv:2310.13673*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large language models are geographically biased. *arXiv preprint arXiv:2402.02680*.

Satish K Mittal. 2022. Behavior biases and investment decision: theoretical and research framework. *Qualitative Research in Financial Markets*, 14(2):213–228.

Mazda Moayeri, Elham Tabassi, and Soheil Feizi. 2024. Worldbench: Quantifying geographic disparities in llm factual recall. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1211–1228.

OpenAI. 2023. Gpt-4 technical report.

Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, and Ziwei Zhu. 2024. Breaking bias, building bridges: Evaluation and mitigation of social biases in llms via contact hypothesis.

David Rozado. 2024. The political preferences of llms. *PloS one*, 19(7):e0306621.

Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, Moritz Roidl, Markus Pauly, et al. 2023. The self-perception and political biases of chatgpt. *Human Behavior and Emerging Technologies*, 2024.

Jayanta Sadhu, Maneesha Rani Saha, and Rifat Shahriyar. 2024. Social bias in large language models for bangla: An empirical study on gender and religious bias. *arXiv preprint arXiv:2407.03536*.

Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768.

Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. 1999. Byte pair encoding: A text compression scheme that accelerates pattern matching.

Carl P Simon, Lawrence Blume, et al. 1994. *Mathematics for economists*, volume 7. Norton New York.

Lars St, Svante Wold, et al. 1989. Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, 6(4):259–272.

Kunsheng Tang, Wenbo Zhou, Jie Zhang, Aishan Liu, Gelei Deng, Shuai Li, Peigui Qi, Weiming Zhang, Tianwei Zhang, and Nenghai Yu. Gendercare: A comprehensive framework for assessing and reducing gender bias in large language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jason Weston and Sainbayar Sukhbaatar. 2023. System 2 attention (is something you might need too).

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023a. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. 2023b. From language modeling to instruction following: Understanding the behavior shift in llms after instruction tuning. *arXiv preprint arXiv:2310.00492*.

Jing Jian Xiao and Nilton Porto. 2019. Present bias and financial behavior. *Financial Planning Review*, 2(2):e1048.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. *arXiv e-prints*, pages arXiv–2309.

Yuhang Zhou, Zeping Li, Siyu Tian, Yuchen Ni, Sen Liu, Guangnan Ye, and Hongfeng Chai. 2024. Silversight: A multi-task chinese financial large language model based on adaptive semantic space learning. *arXiv preprint arXiv:2404.04949*.

A Cognitive Bias

For cognitive bias, we classified it into Belief Bias and Risk reference Bias based on previous research, and studied seven of these biases within the FBI framework.Refer to Table 2 for specific content.

Table 2: Summary of Cognitive Biases

Cognitive Bias	Bias Type	Definition
Belief Bias	Limited Attention	The brain has two systems when working: fast thinking and slow thinking. It uses intuition to deal with things quickly.
	Representativeness bias	When making probability estimates, people tend to focus on certain representative features, ignoring environmental probabilities and sample size.
	Anchoring effect	Decision-making is often influenced by the first information received, like an anchor sinking to the bottom of the sea.
	Overconfidence	Belief that one's knowledge is more accurate than the facts; one's information is given more weight.
Risk-Preference Bias	Situational dependence bias	The effect of a stimulus depends largely on the context in which it occurs.
	Loss aversion	Sensitivity to losses exceeds gains of equal value.
	Framing effect	Different descriptions of an objectively identical problem lead to different decision-making judgments.

B Company Profile

In order to avoid bias caused by market value impact, we did not choose funds from the CSI 300 or CSI 500. Instead, we summarized all listed companies in China. After removing ST type stocks, we selected the top, middle, and bottom 200 stocks based on market value, totaling 600 stocks. The industry distribution of stocks is shown in the Figure 8.

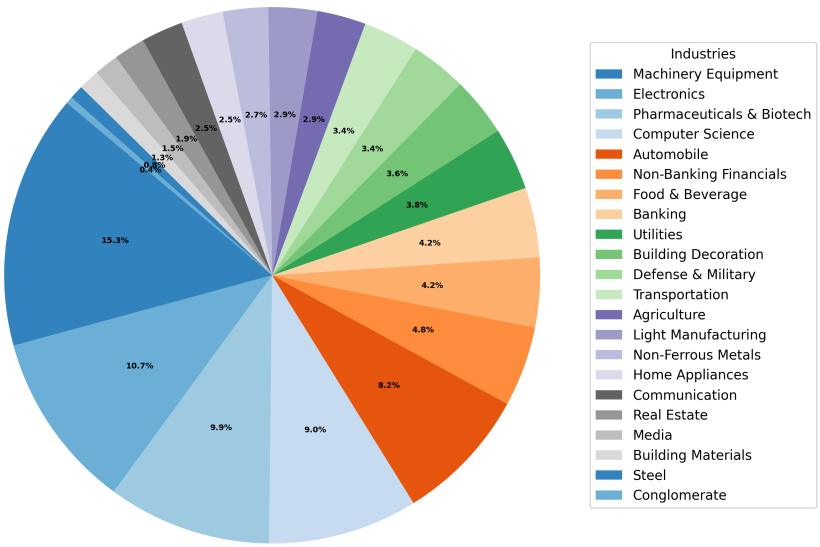


Figure 8: Distribution of the selectd company's industry type.

C Event Type

We have sorted out the types of events that can affect a company's stock price based on the regular patterns of the Chinese A-Share stock market, and finally sorted out four categories, totaling 16 types of events. The detailed content is shown in Table 3.

D Models

We have selected a total of 23 financial and general LLMs oriented by Chinese and English, with specific details shown in Table 4.

E Formula Proof

Invoking the fundamental principles of expected utility theory, we recognize that a utility function's curvature reflects an individual's risk preference. Specifically, a concave utility function ($u''(x) < 0$) is indicative of risk aversion, while a convex utility function ($u''(x) > 0$) signifies risk-seeking behavior. A linear utility function ($u''(x) = 0$), on the other hand, corresponds to risk neutrality.

The expected utility $E[u(x)]$ can be formally represented as:

$$E[u(x)] = \sum_{x(\omega)} u(x(\omega))p(x(\omega)) \tag{5}$$

Here, $u(x)$ denotes the utility function, $x(\omega)$ symbolizes the outcome under state ω, and $p(x(\omega))$ is the probability of outcome $x(\omega)$ occurring.

Furthermore, we articulate the variance of outcomes x, $\text{Var}(x)$, as the expected squared deviation from the expected value $E[x]$:

$$\text{Var}(x) = E[(x - E[x])^2] \tag{6}$$

Applying the second-order Taylor expansion to the utility function $u(x)$ around the expected value $E[x]$ furnishes us with:

$$u(x) \approx u(E[x]) + u'(E[x])(x - E[x]) + \frac{1}{2}u''(E[x])(x - E[x])^2 \tag{7}$$

Imposing expectations on the approximated function, we derive the expected utility approximation:

$$E[u(x)] \approx u(E[x]) + u'(E[x])E[x - E[x]] + \frac{1}{2}u''(E[x])E[(x - E[x])^2] \tag{8}$$

Since $E[x - E[x]] = 0$, the middle term vanishes, simplifying our expression to:

$$E[u(x)] \approx u(E[x]) + \frac{1}{2}u''(E[x])\text{Var}(x) \tag{9}$$

Consequently, under the assertion of utility function concavity or convexity, the sign of the second derivative $u''(E[x])$ establishes the nature of risk preference. For a negative second derivative ($u''(x) < 0$), indicative of risk aversion, a smaller variance is required to enhance the expected utility. Conversely, for a positive second derivative ($u''(x) > 0$), characteristic of risk-seeking behavior, a larger variance is preferred. Risk-neutral individuals ($u''(x) = 0$) show indifference to the variance level.

Through this analytical framework, we delineate how the variance of outcomes in conjunction with the utility function's concavity or convexity guides the determination of an individual's risk preference.

F Framework for Data Construction

This section delineates the structured approach employed in the study to formulate datasets incorporating event news, interactive elements, and risk preference inquiries. Each category of information is meticulously crafted using a distinct template, which is elucidated below.

Table 3: Event Types and Definitions

Event Type	Subdivision Type	Definition
Corporate Governance and Equity Changes	Major Asset Restructuring	The process of recombining, adjusting, and allocating the distribution status of enterprise assets among the owners, controllers, and external economic entities.
	Equity Incentive	By conditionally granting employees partial shareholder rights, a sense of ownership is fostered, forming a community of interests with the company.
	Increase or Decrease in Shareholder Holdings	Changes in the shareholder holdings of company stocks.
	Buy-back	The act of a listed company using cash or other means to repurchase its shares from the stock market.
	Circulation of Restricted Stock	Restricted shares become freely tradable in the secondary market after the commitment period.
Financial Reports and Earnings Expectations	Performance Report	Regular preparation by each responsibility center to evaluate and assess performance, serving as the basis for future budget preparation.
Market Behavior and Announcements	Private Placement	Targeted issuance of bonds or stocks to a select group of senior institutional or individual investors.
	Transfer of Shares	Listed companies transfer their provident fund to share capital in proportion or issue bonus shares accordingly.
	Stock Price Fluctuations	Sudden large inflows and outflows of funds lead to increased volatility in stock prices.
	Business Dynamics	Updates on enterprises and their surroundings, using major production and sales information to promote corporate brand and image.
Negative Events and Risk Management	Dispute	Disputes between companies or between companies and individuals.
	Investigation	Filing an investigation signifies a basic determination of illegal facts, allowing for compulsory measures and official initiation of investigation procedures.
	Violation Penalties	Punishments for enterprises violating regulations of regulatory bodies.
	Litigation and Arbitration	Litigation and arbitration for contract disputes and other property rights disputes between enterprises.
	Security	Enterprises providing guarantees for loans and other matters for other enterprises.

Table 4: Models in our Framework

Model Name	**Chinese-oriented**	**Model size**	**FinLLM**	**Deployment method**
MiniCPM-2B	True	2B	False	local
Baichuan-13B	True	13B	False	local
DISC-FinLLM	True	13B	True	local
Baichuan2-7B	True	7B	False	local
Baichuan2-13B	True	13B	False	local
ChatGLM2-6B	True	6B	False	local
ChatGLM3-6B	True	6B	False	local
ChatGLM3-Turbo	True	33B	False	API
GLM-4	True	Unknown	False	API
InternLM2-7B	True	7B	False	local
InternLM2-20B	True	20B	False	local
LLaMA2-7B	False	7B	False	local
LLaMA2-13B	False	13B	False	local
Qwen-7B	True	7B	False	local
Qwen-14B	True	14B	False	local
FinQwen	True	14B	True	local
Qwen-72B	True	72B	False	local
Qwen-max	True	72B	False	API
Xuanyuan-13B	True	13B	True	local
Xuanyuan-70B	True	70B	True	local
Xuanyuan2-6B	True	6B	True	local
GPT-3.5	False	Unknown	False	API
GPT-4	False	Unknown	False	API

F.1 Event News Template

The construction of the event news dataset leverages prompt engineering techniques to embed real-world events within a framework that facilitates evaluation, simulating the analytical capabilities of financial experts. The evaluation process involves the model assigning a score to each event based on its potential positive or negative impact on the financial landscape. Initially, the model is instructed to provide an immediate, intuitive score reflecting a 'fast thinking' approach.

To augment the depth of analysis and ensure the robustness of the evaluation, the model is further tasked with adopting a 'slow thinking' strategy. This entails a comprehensive articulation of the rationale behind the score, encouraging a deliberate and reasoned assessment. The detailed format of this template is illustrated in Figure 9, which guides the model in delivering both the quantitative score and the qualitative reasoning underpinning it.

事件： 公司X：主业强势升级可期 买入评级　　公司X于2023年2月3日发布公告：扩增行业Y相关经营管理，在保证主营业务不变的情况下拓展业务范围。

Event: *Company X*: Strong Upgrade of Main Business Expected Purchase Rating *Company X* announced on February 3, 2023: Expanding *Industry Y* related business management and expanding business scope while ensuring the main business remains unchanged.

提示：请忘记之前的指令。假装自己是一位金融专家，将要对以下新闻进行打分，利好为正分，利空为负分，分数区间为-10到+10分，要求直接输出分数而不输出多余的解释等内容。
以下将贵州茅台酒股份有限公司简称为贵州茅台，请对其进行打分。
贵州茅台：主业强势升级可期 买入评级　　贵州茅台于2023年2月3日发布公告：扩增食品饮料相关经营管理，在保证主营业务不变的情况下拓展业务范围。

Instruct: Please forget the previous instructions. Pretend to be a financial expert and score the following news with positive scores for positive and negative scores for negative, ranging from -10 to+10 points. Require direct output of scores without unnecessary explanations. Hereinafter, *Kweichow Moutai Liquor Co., Ltd.* is referred to as Kweichow Moutai, please give it a score. *Kweichow Moutai*: The main business of *Kweichow Moutai* will upgrade its buying rating strongly. On February 3, 2023, it will issue an announcement: expand the operation and management related to *Food & Beverage*, and expand its business scope while ensuring that the main business remains unchanged.

思维链：请忘记之前的指令。假装自己是一位金融专家，将要对以下新闻进行打分，利好为正分，利空为负分，分数区间为-10到+10分，请在输出分数前先输出100字内的分析，再换行后输出分数。
以下将贵州茅台酒股份有限公司简称为贵州茅台，请对其进行打分。
贵州茅台：主业强势升级可期 买入评级　　贵州茅台于2023年2月3日发布公告：扩增食品饮料相关经营管理，在保证主营业务不变的情况下拓展业务范围。

COT: Please forget the previous instructions. Pretend to be a financial expert and score the following news with positive scores for positive and negative scores for negative, ranging from -10 to+10 points. Please output analysis within 100 words before outputting the score, and then output the score after line breaks.
Hereinafter, *Kweichow Moutai Liquor Co., Ltd.* is referred to as Kweichow Moutai, please give it a score. *Kweichow Moutai*: The main business of *Kweichow Moutai* will upgrade its buying rating strongly. On February 3, 2023, it will issue an announcement: expand the operation and management related to *Food & Beverage*, and expand its business scope while ensuring that the main business remains unchanged.

Figure 9: Template of event news.

F.2 Interactions

Using input methods similar to news events for rewriting, the specific template is shown in Figure 10.

交互：
问题:请问截至1月31日收盘公司股东人数是多少?
回答：您好，截至2024年1月31日公司X的股东人数为22,554。

Interaction:
Question: May I ask what is the number of shareholders in the company as of the close of January 31st?
Answer: Hello, as of January 31, 2024, the number of shareholders of *Company X* is 22554.

指令提示：请忘记之前的指令。假装自己是一位金融专家，将要对以下公司回股民问进行打分，利好为正分，利空为负分，分数区间为-10到+10分，仅对公司回复进行判断，要求直接输出分数而不输出多余的解释等内容。以下将宁德时代新能源科技股份有限公司简称为宁德时代，且下文为其回股民问，请对回复进行打分。
问题:请问截至1月31日收盘公司股东人数是多少?
回答：您好，截至2024年1月31日宁德时代公司股东人数为22,554。

Instruct: Please forget the previous instructions. Pretending to be a financial expert, I am going to score the following company's stock investors with positive scores for positive and negative scores for negative, ranging from -10 to+10 points. I will only judge the company's response and require direct output of scores without unnecessary explanations. Below, *Ningde Times New Energy Technology Co., Ltd.* is referred to as *Ningde Times*, and the following is a question from its shareholders. Please rate the response.
Question: May I ask what is the number of shareholders in the company as of the close of January 31st?
Answer: Hello, as of January 31, 2024, the number of shareholders of *Ningde Times* Company is 22554.

思维链：请忘记之前的指令。假装自己是一位金融专家，将要对以下公司回股民问进行打分，利好为正分，利空为负分，分数区间为-10到+10分，仅对公司回复进行判断，请在输出分数前先输出100字内的分析，再换行后输出分数。以下将宁德时代新能源科技股份有限公司简称为宁德时代，且下文为其回股民问，请对回复进行打分。
问题:请问截至1月31日收盘公司股东人数是多少?
回答：您好，截至2024年1月31日宁德时代公司股东人数为22,554。

COT: Please forget the previous instructions. Pretending to be a financial expert, I am going to score the following company's stock investors with positive scores for positive and negative scores for negative, ranging from -10 to+10 points. I will only judge the company's response. Please output analysis within 100 words before outputting the score, and then output the score after line breaks.Below, *Ningde Times New Energy Technology Co., Ltd.* is referred to as *Ningde Times*, and the following is a question from its shareholders. Please rate the response.
Question: May I ask what is the number of shareholders in the company as of the close of January 31st?
Answer: Hello, as of January 31, 2024, the number of shareholders of *Ningde Times* Company is 22554.

Figure 10: Template of interactions.

F.3 Risk-Preference Questionnaire Template

The methodology for assessing risk preferences through structured questions is twofold, designed to discern the inherent risk orientation of the AI model under different conditions. Initially, the model is presented with a set of scenarios where it must select an option that best aligns with its assessed risk profile, simulating an introspective decision-making process. This setup aims to capture the model's spontaneous risk preferences without external biases.

Subsequently, the experiment introduces a predefined constraint by explicitly characterizing the model as risk-averse within the instructions. This manipulation is intended to observe the adaptability of the model's responses to altered risk parameters, thereby evaluating its capacity for contextual behavioral adjustment. The layout and content of these questions are encapsulated in the template depicted in Figure 11, which systematically guides the model through the decision-making process under varying risk conditions.

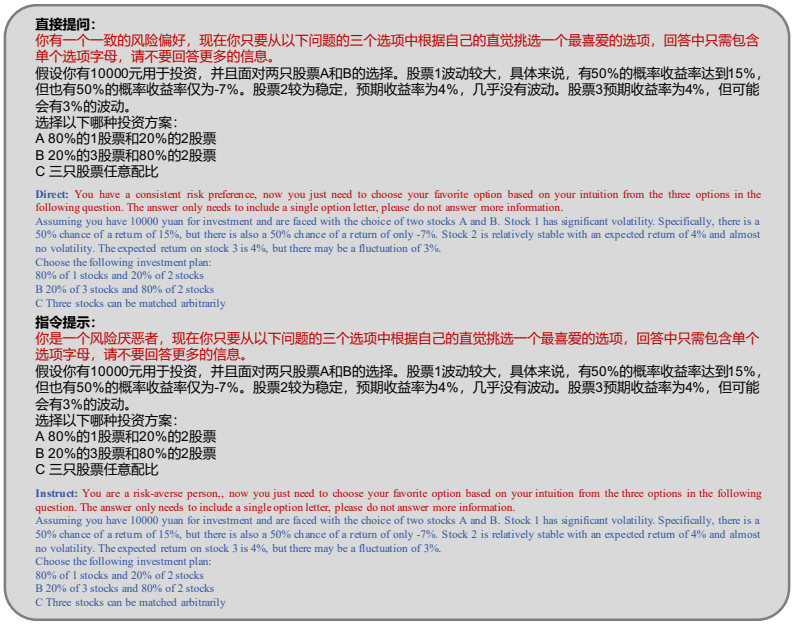


Figure 11: Template of risk-preference questions.

G FinCausal Dataset

In this section, we introduce the construction process of the FinCausal dataset, which requires the acquisition of relevant causal knowledge from past financial text materials. This process can be mainly divided into data collection, deduplication, segmentation, and knowledge extraction. The main flowchart is shown in Figure 12.

Firstly, we crawled 500,000 research reports from the internet, ranging from 2020 to 2023, including individual stock research, industry analysis, and macro analysis. We used regular matching and the FastText language filter for classification, mainly retaining Chinese A-Share individual stock research and industry analysis. Since individual stock reports rarely describe causal relationships for negative events, we further crawled some news analyses and stock forum comments to enrich the description of individual stock causal knowledge.

After filtering the research report data, we used the MinHash algorithm to perform document-level deduplication on all content. To extract sentences with causal expressions, we meticulously categorized the content of the research reports into seven distinct types, which included ordinary sentences, causal sentences, news-related content, recommendation ratings, investment advice, risk warnings, and researcher information. We then manually annotated a comprehensive dataset comprising 3,000 pieces of data to train a sophisticated BGE+TextCNN classification model. This model was specifically designed to discern

and categorize the various types of sentences present in the financial reports, with a particular focus on identifying those that convey causal relationships.

For each piece of extracted knowledge from a research report or comment, we concatenated one sentence before and after it into a paragraph. Subsequently, we aggregated all relevant paragraphs from a report to form a comprehensive context. Utilizing our meticulously chosen LLM, we conducted causal knowledge extraction from these contexts. Through this process, we successfully obtained 200,000 pieces of industry causal knowledge and 2,000 pieces of individual stock causal knowledge, thereby enriching our dataset with valuable insights into the causal relationships within the financial domain. Here are some examples of FinCausal:

- The company may consider conducting targeted issuance in order to expand its business scale, make capital expenditures, or invest in research and development.
- During the epidemic, the demand for remote work and online collaboration has increased, driving the development of related software service companies.

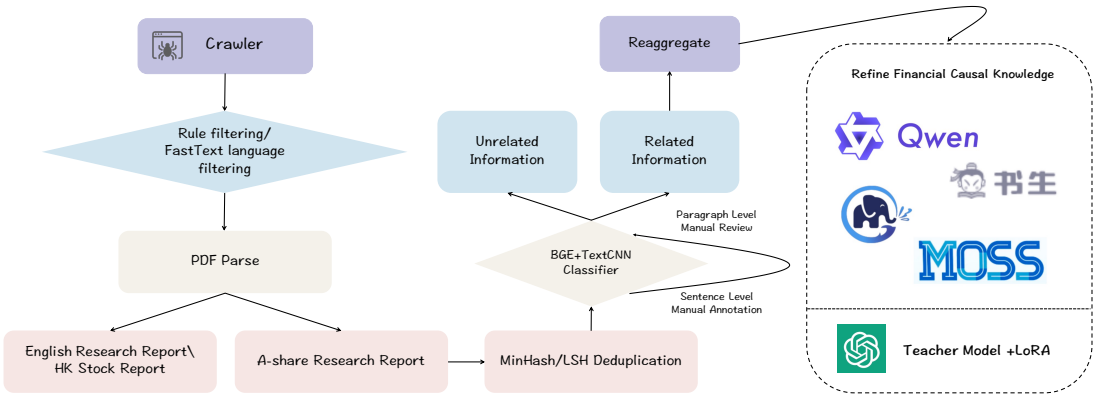


Figure 12: The construction process of FinCausal dataset.

H Result

H.1 Analysis of Direct News Events

The examination of news events involves a detailed statistical analysis of the responses generated by various Large Language Models (LLMs) to specific news items. This analysis primarily focuses on the distribution of scores assigned by LLMs to each news event, encompassing key statistical measures such as the mean, variance, highest, and lowest scores. Such an approach is instrumental in assessing the consistency and rationality of LLMs' interpretations of financial news.

To facilitate a comprehensive understanding of these scoring distributions, this section will present violin plots for each news event. Violin plots offer a more nuanced visualization compared to traditional box plots by showing the probability density of the data at different values. This graphical representation will thus provide insights into the spread and skewness of LLMs' ratings across various news events, enabling a deeper analysis of their evaluative patterns and potential biases.

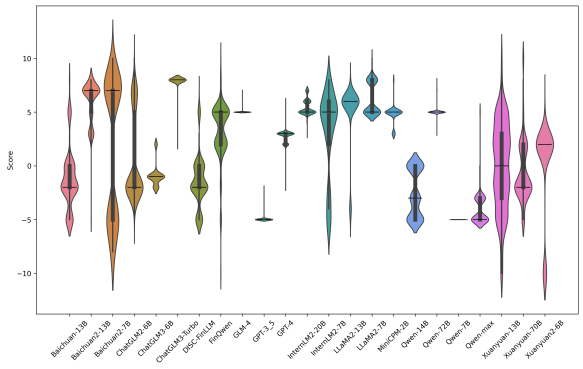


Figure 13: Distribution of the score for news 1 among 23 large models.

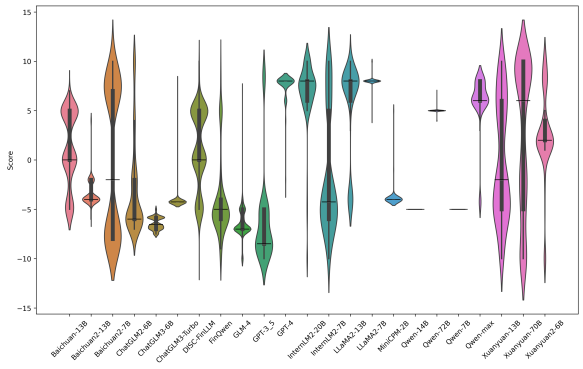


Figure 14: Distribution of the score for news 2 among 23 large models.

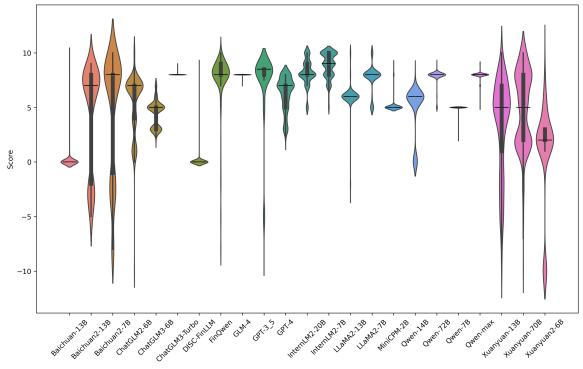


Figure 15: Distribution of the score for news 3 among 23 large models.

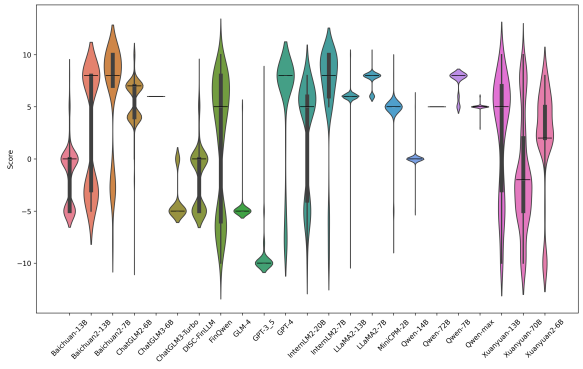


Figure 16: Distribution of the score for news 4 among 23 large models.

In the analysis of the initial five events, as shown in Figure 13 to Figure 36 the focus is placed on news items that encompass both positive and negative performance reports, alongside fluctuations in stock prices. This diverse array of news content allows for a multifaceted examination of each Large Language Model's (LLM's) scoring tendencies. Notably, discrepancies in scoring preferences among different LLMs emerge when confronted with this spectrum of financial news.

A systematic statistical analysis is conducted on the scoring outcomes attributed to the positive and negative aspects of these events. This entails a detailed examination of how each LLM assesses the same news piece, shedding light on the variance in their interpretations and the potential implications of their biases. The findings from this analysis are meticulously compiled and presented in Table 5, offering a clear, quantified insight into the LLMs' evaluative patterns across the selected news events.

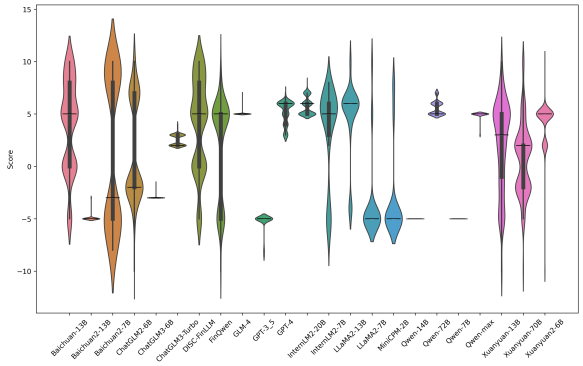


Figure 17: Distribution of the score for news 5 among 23 large models.

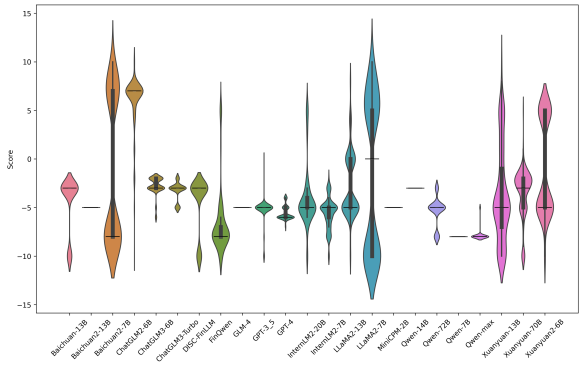


Figure 18: Distribution of the score for news 6 among 23 large models.

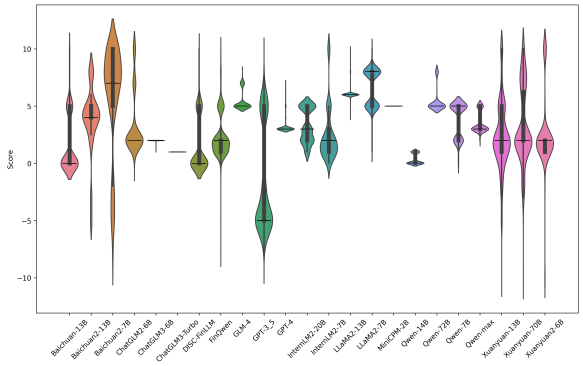


Figure 19: Distribution of the score for news 7 among 23 large models.

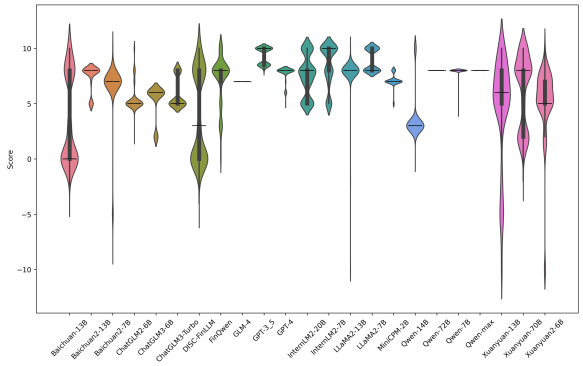


Figure 20: Distribution of the score for news 8 among 23 large models.

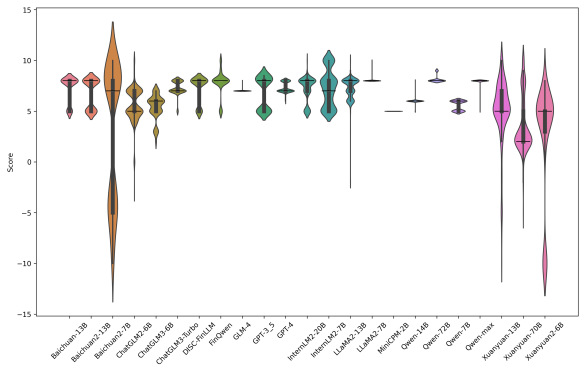


Figure 21: Distribution of the score for news 9 among 23 large models.

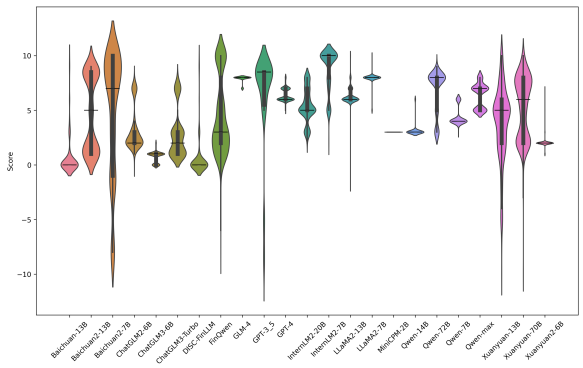


Figure 22: Distribution of the score for news 10 among 23 large models.

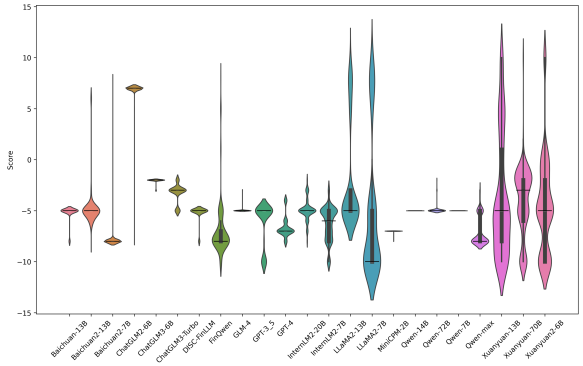


Figure 23: Distribution of the score for news 11 among 23 large models.

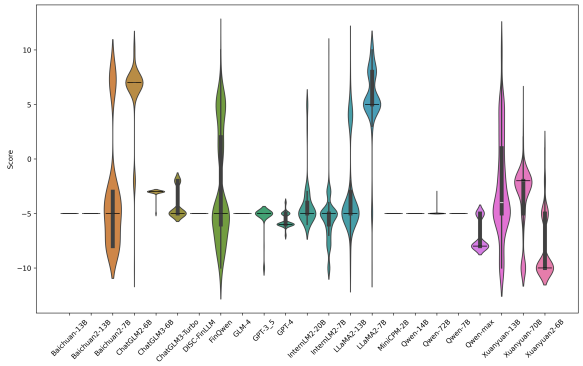


Figure 24: Distribution of the score for news 12 among 23 large models.

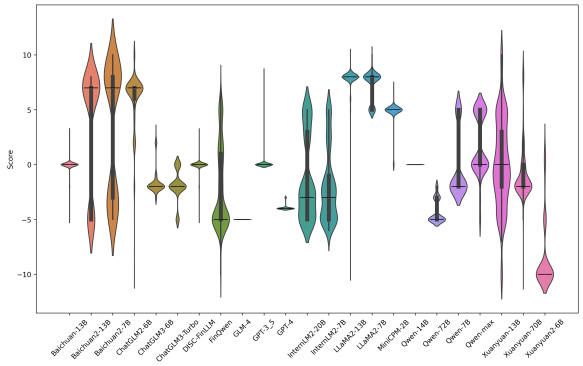


Figure 25: Distribution of the score for news 13 among 23 large models.

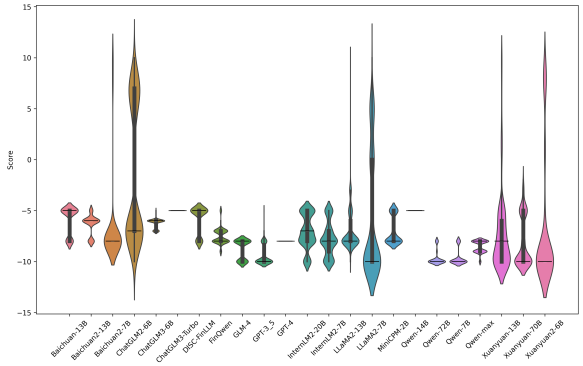


Figure 26: Distribution of the score for news 14 among 23 large models.

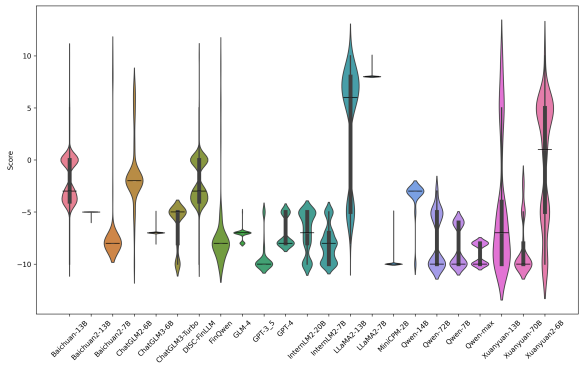


Figure 27: Distribution of the score for news 15 among 23 large models.

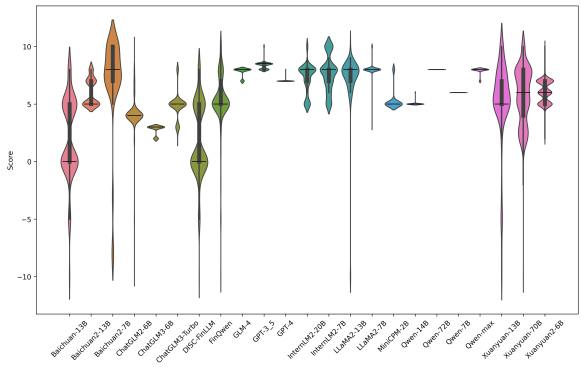


Figure 28: Distribution of the score for news 16 among 23 large models.

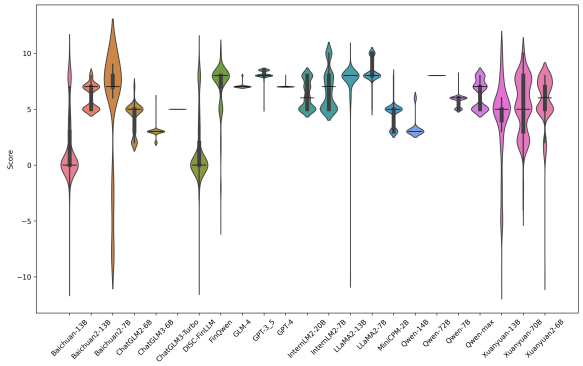


Figure 29: Distribution of the score for news 17 among 23 large models.

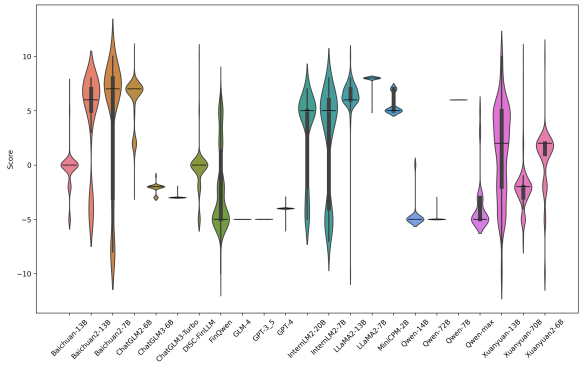


Figure 30: Distribution of the score for news 18 among 23 large models.

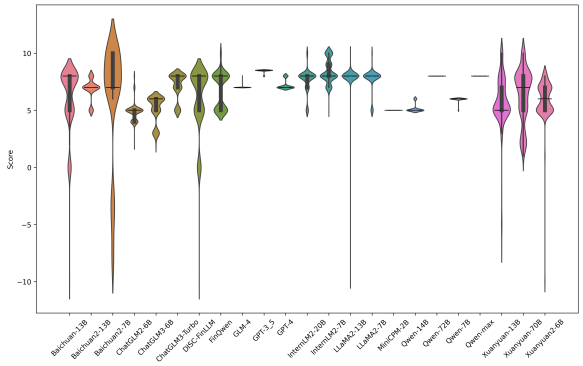


Figure 31: Distribution of the score for news 19 among 23 large models.

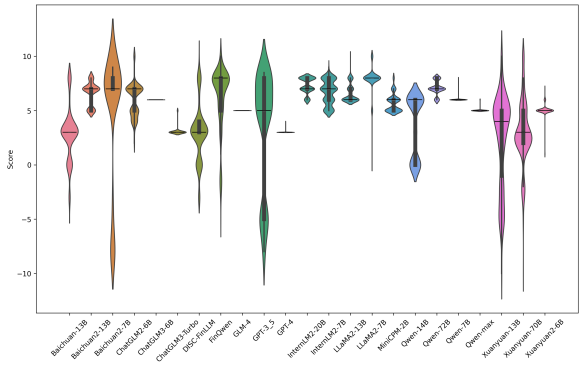


Figure 32: Distribution of the score for news 20 among 23 large models.

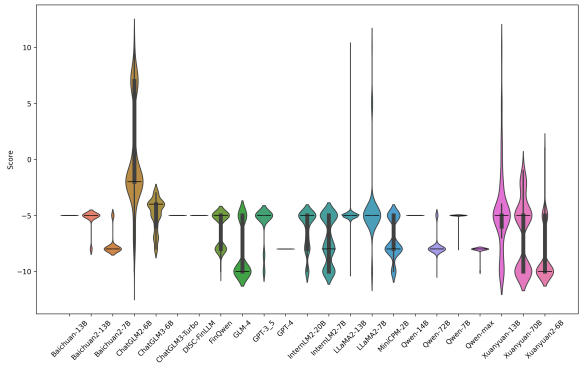


Figure 33: Distribution of the score for news 21 among 23 large models.

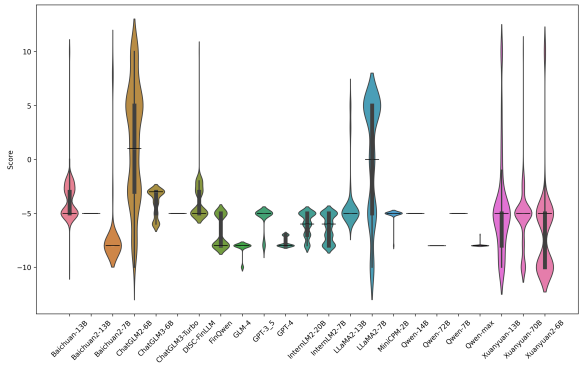


Figure 34: Distribution of the score for news 22 among 23 large models.

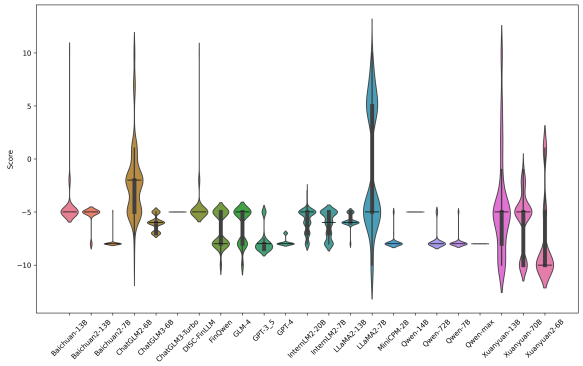


Figure 35: Distribution of the score for news 23 among 23 large models.

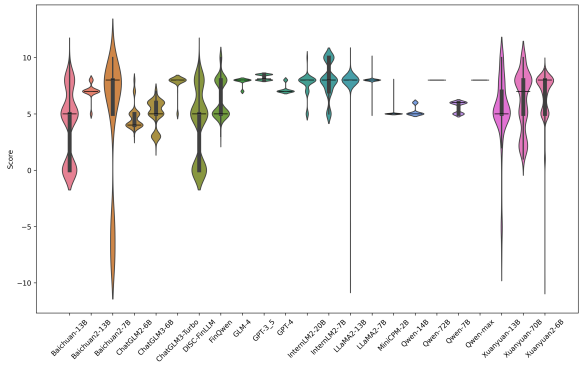


Figure 36: Distribution of the score for news 24 among 23 large models.

Table 5: Model Positive Times

Model	**Positive Times**
GPT-4	5
InternLM2-20B	5
LLaMA2-13B	5
Qwen-72B	5
FinQwen	4
InternLM2-7B	4
LLaMA2-7B	4
Qwen-max	4
Xuanyuan-13B	4
Baichuan2-13B	3
Baichuan2-7B	3
ChatGLM3-Turbo	3
GLM-4	3
Xuanyuan-70B	3
ChatGLM2-6B	2
ChatGLM3-6B	2
Qwen-14B	2
Qwen-7B	2
GPT-3.5	1

Our analysis involves aggregating the scoring variances observed across all event news for the various Large Language Models (LLMs) under consideration. This comprehensive synthesis not only highlights the diversity in LLM responses but also provides a macroscopic view of their evaluative consistency and potential discrepancies. The aggregated data, which encapsulate the variance in scoring for each news event by different LLMs, are systematically presented in Table 6. This table serves as a pivotal reference point for understanding the range and distribution of LLM evaluations, offering valuable insights into their interpretative frameworks and the reliability of their analyses.

Table 6: Model Variance Comparison

Model	**Variance**
GLM-4	0.59798884
ChatGLM3-6B	0.707638507
Qwen-72B	0.784471341
Qwen-7B	0.788077699
ChatGLM3-Turbo	1.067120654
Qwen-14B	1.211226324
Qwen-max	1.363195393
MiniCPM-2B	1.409
GPT-4	1.909388332
InternLM2-20B	4.893324616
GPT-3.5	5.277003518
Baichuan-13B	5.998
DISC-FinLLM	6.096
Baichuan2-13B	6.157081681
LLaMA2-13B	6.881628177
InternLM2-7B	7.466014898
FinQwen	9.363445905
ChatGLM2-6B	10.03005785
LLaMA2-7B	10.61274671
Xuanyuan-70B	10.83743438
Xuanyuan2-6B	13.9988
Xuanyuan-13B	19.18007393
Baichuan2-7B	28.10579705

Upon examining the inherent biases within individual models, our analysis proceeds to consolidate the findings from each Large Language Model (LLM) to explore their collective or differential biases towards various industries. This step is crucial for understanding not only the predispositions of individual models but also for discerning any overarching trends or anomalies in their assessments of industry-related news events. By aggregating these results, we aim to delineate the extent to which these models exhibit preferential or adverse biases towards certain industry, thereby shedding light on the potential influence of these biases on the models' analytical outputs and reliability. The synthesis of this comprehensive analysis provides a nuanced understanding of model behavior in the context of industry-specific evaluations.

In parallel with the examination of model biases, our study also delves into the temporal evolution of Large Language Models within distinct family, attributing changes to factors such as model size or software updates. To this end, we employ box line comparison charts as a visual tool to elucidate the developmental trajectories of models within the Baichuan, GLM, and LLaMA family. These charts serve to highlight variations in model performance or bias over time, providing a clear visual representation of progression or shifts in model behavior. The comparative analyses for the Baichuan family, GLM family, and LLaMA family are depicted in Figure 56, Figure 57, and Figure 58, respectively. Through these visual comparisons, we aim to offer insights into how advancements or modifications in model architecture and capabilities influence their analytical outcomes and biases.

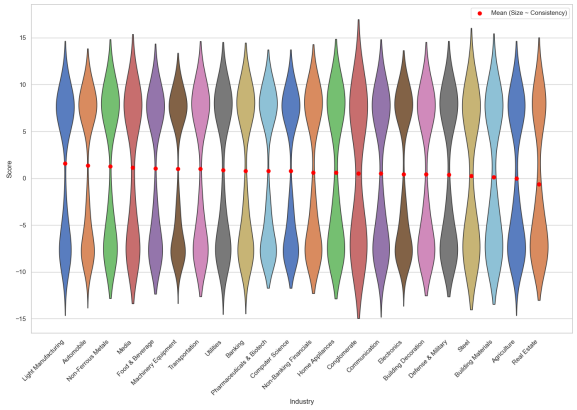


Figure 37: Distribution of the industry scores of Baichuan2-7B.

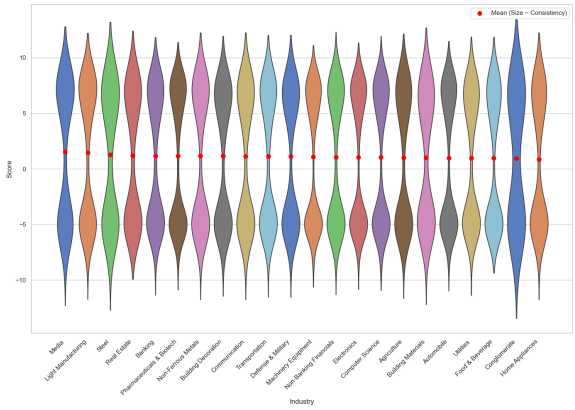


Figure 38: Distribution of the industry scores of Baichuan2-13B.

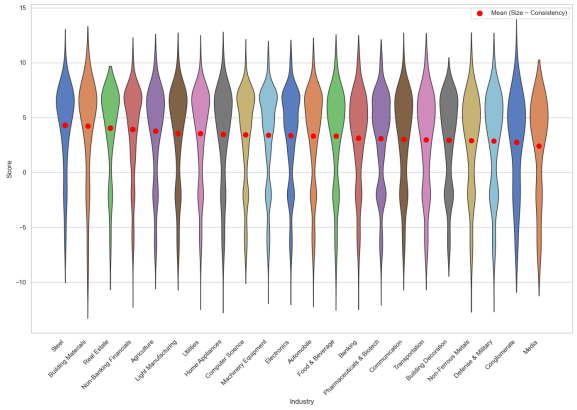


Figure 39: Distribution of the industry scores of ChatGLM2-6B.

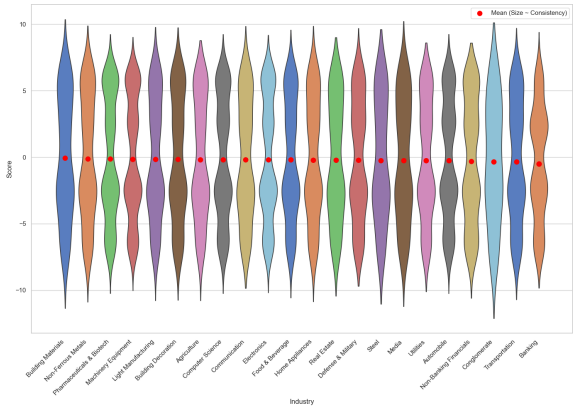


Figure 40: Distribution of the industry scores of ChatGLM3-6B.

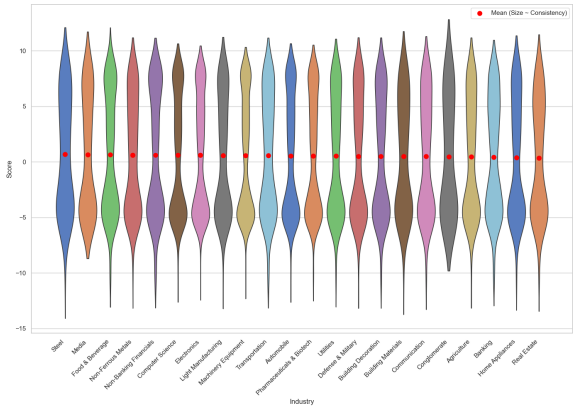


Figure 41: Distribution of the industry scores of ChatGLM3-Turbo.

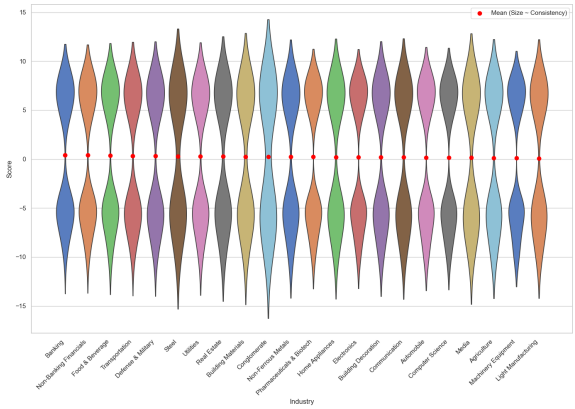


Figure 42: Distribution of the industry scores of GLM-4.

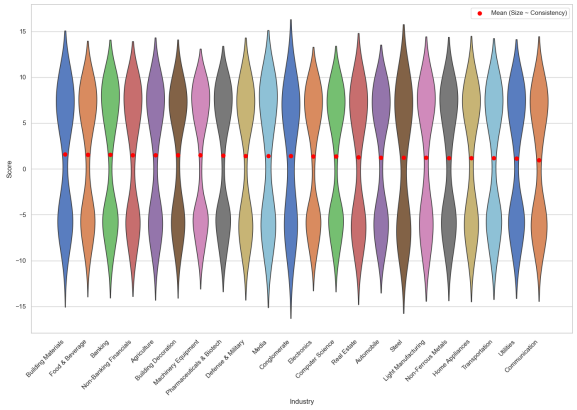


Figure 43: Distribution of the industry scores of InternLM2-7B.

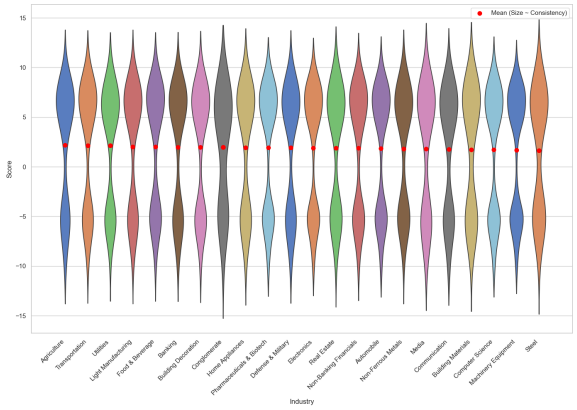


Figure 44: Distribution of the industry scores of InternLM2-20B.

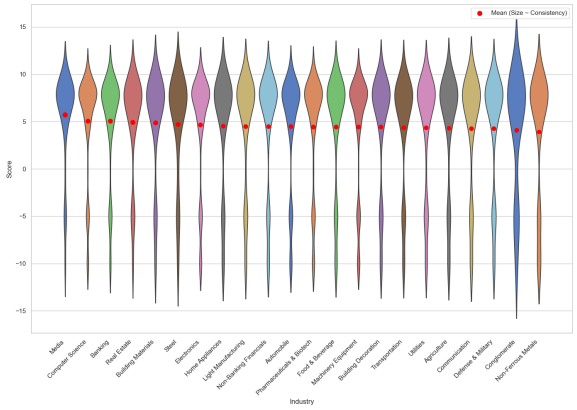


Figure 45: Distribution of the industry scores of LLaMA2-7B.

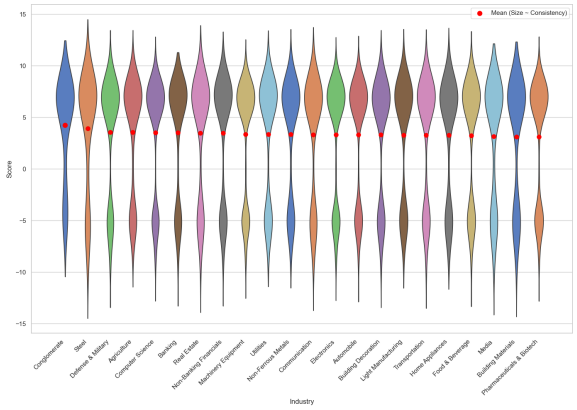


Figure 46: Distribution of the industry scores of LLaMA2-13B.

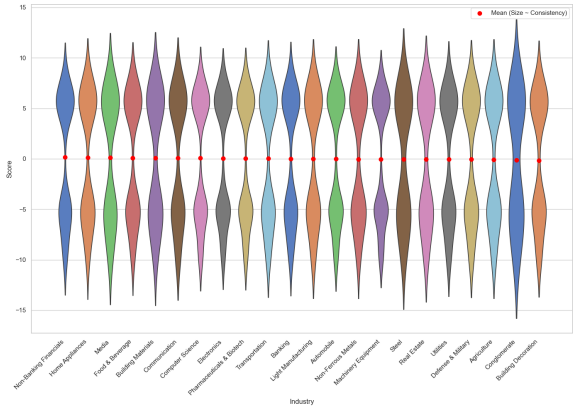


Figure 47: Distribution of the industry scores of Qwen-7B.

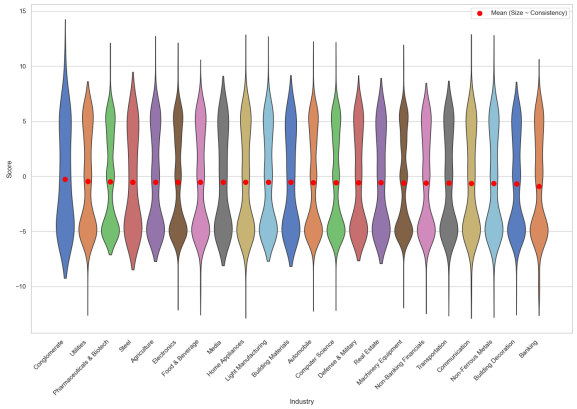


Figure 48: Distribution of the industry scores of Qwen-14B.

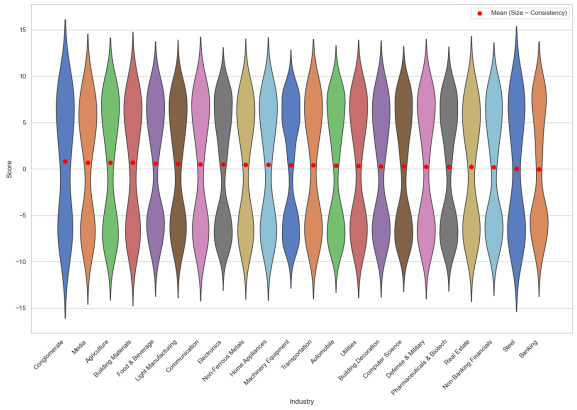


Figure 49: Distribution of the industry scores of Fin-Qwen.

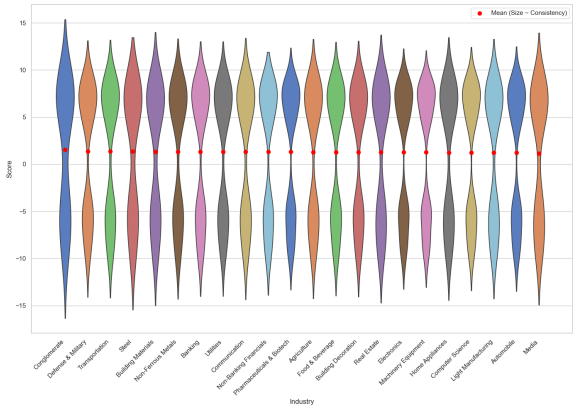


Figure 50: Distribution of the industry scores of Qwen-72B.

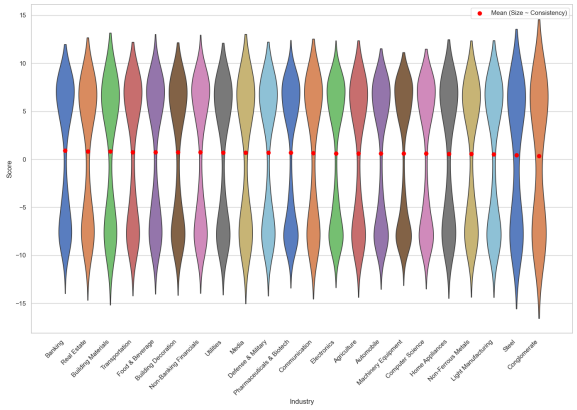


Figure 51: Distribution of the industry scores of Qwen-max.

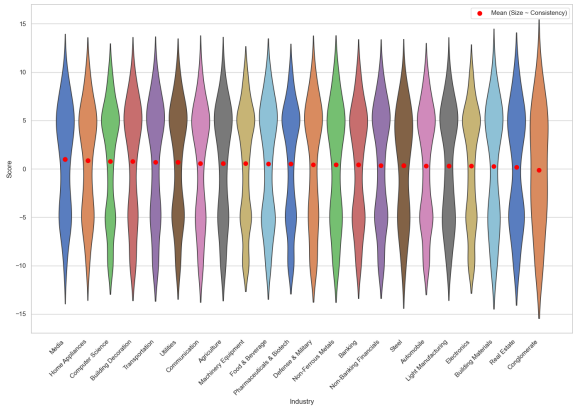


Figure 52: Distribution of the industry scores of Xuanyuan-13B.

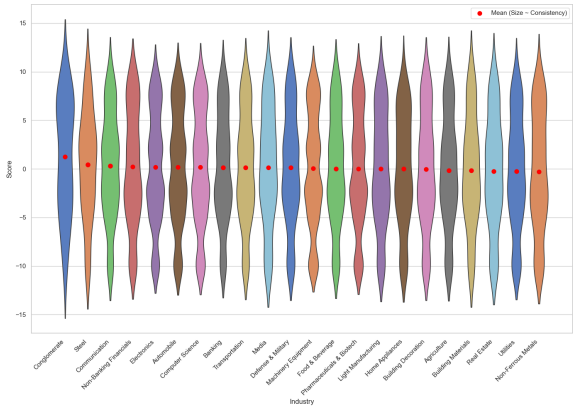


Figure 53: Distribution of the industry scores of Xuanyuan-70B.

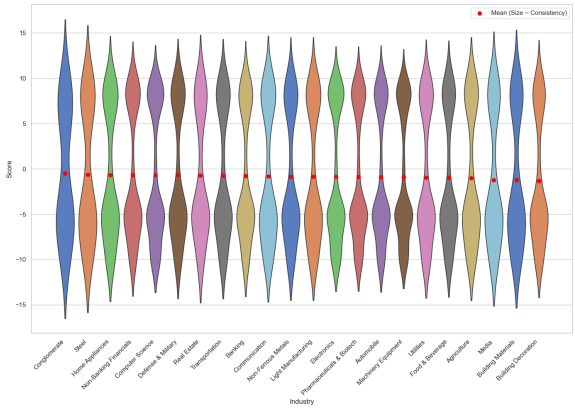


Figure 54: Distribution of the industry scores of GPT-3.5.

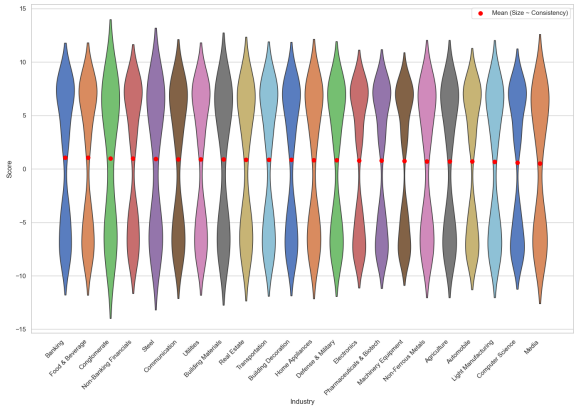


Figure 55: Distribution of the industry scores of GPT-4.

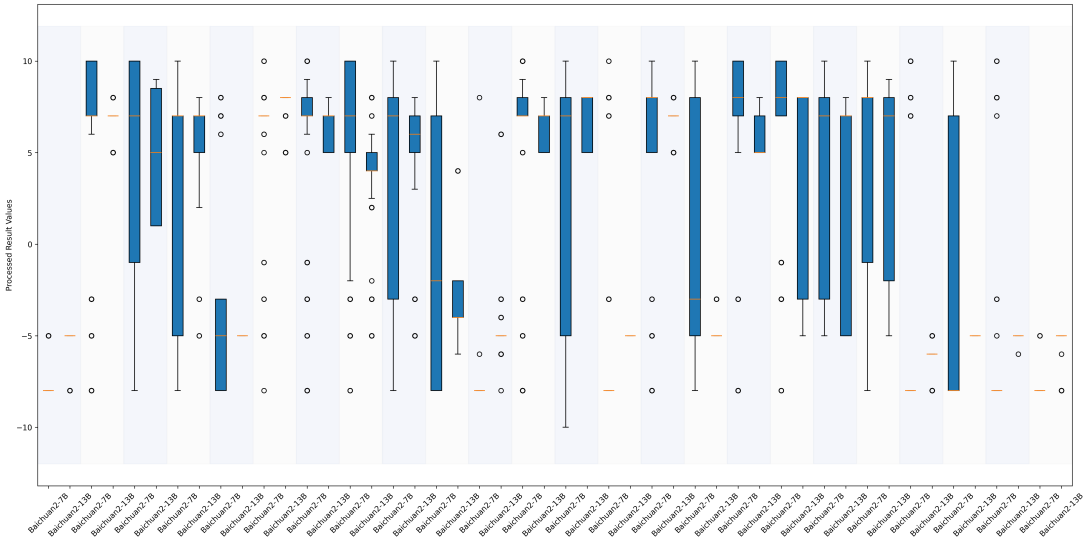


Figure 56: Box line comparison charts of Baichuan family.

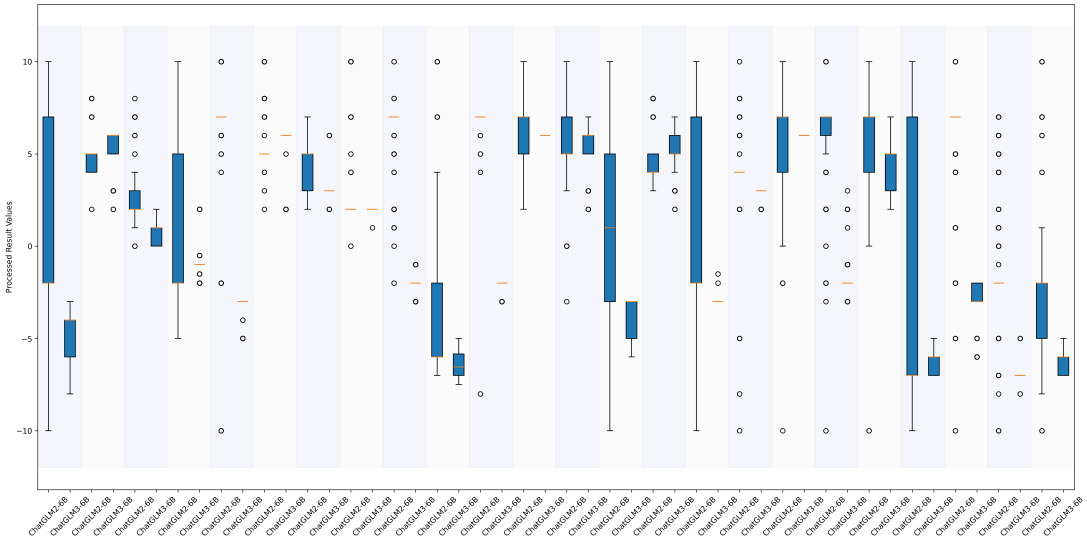


Figure 57: Box line comparison charts of GLM family.

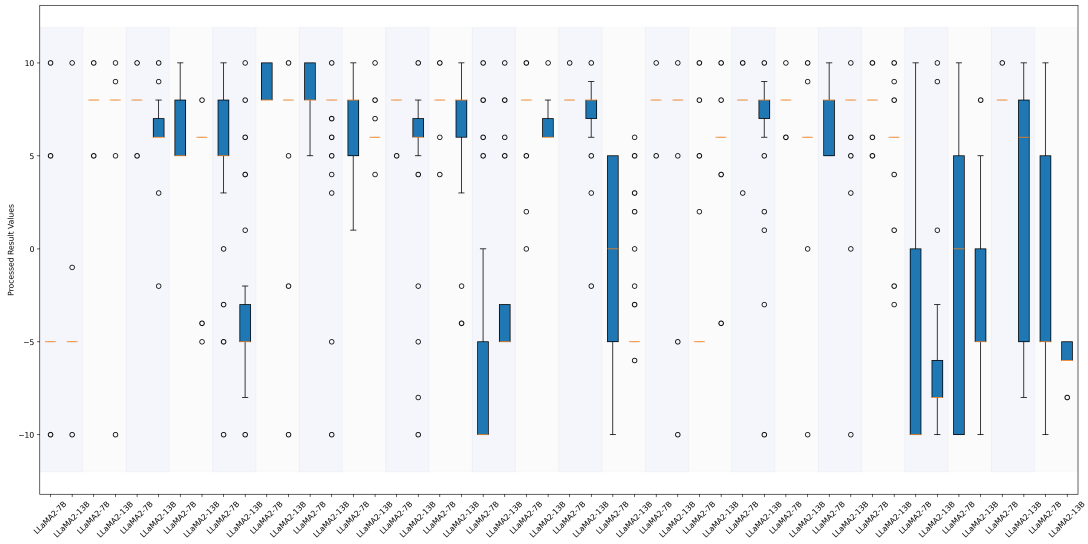


Figure 58: Box line comparison charts of LLaMA family.

H.2 Analysis of COT News

To delve deeper into the underlying factors contributing to potential irrationalities in Large Language Models (LLMs), our investigation extends to the analysis of reasoning outcomes facilitated by cognitive connections. This approach is predicated on the hypothesis that the manner in which LLMs forge and utilize cognitive links during the reasoning process may shed light on their logical inconsistencies or biases. For this purpose, we have meticulously selected LLMs that have demonstrated the highest, second highest, and lowest levels of performance in response to direct prompts. This selection criterion ensures a comprehensive overview, encompassing a broad spectrum of reasoning capabilities within LLMs.

The focus of this analysis is on the 'slow thinking' aspect of model reasoning, where deliberate and methodical processing is emphasized. By examining the variance in reasoning outcomes among these models, we aim to identify patterns or anomalies that might indicate a propensity for irrational decision-making. The results of this analysis, highlighting the variance in cognitive reasoning among the selected LLMs, are systematically presented in Table 7. Through this examination, we seek to uncover the intricacies of cognitive processing in LLMs and their implications for model reliability and rationality.

Table 7: Model Variance Comparison after COT

Model	Direct	COT
GLM-4	0.597988840	5.381799977
Qwen-7B	0.788077699	7.685484073
ChatGLM3-Turbo	1.067120654	5.670563704
MiniCPM-2B	1.409476674	7.038961848
Xuanyuan2-6B	13.99883011	6.668694967
Xuanyuan-13B	19.18007393	17.83545943
Baichuan2-7B	28.10579705	12.65975644

At the same time, we further analyzed the new5 with significant differences in ratings among different LLMs, and used the keyword detection method Bertopic to cluster and analyze the reasoning results of the models. Before clustering, the model scores and specific information of the company were removed from the reasoning results. The inference decibel of each model is clustered into 10 categories, and the distribution of scores for each category is as follows.

We further analyze the reasoning texts of the best performing GLM-4 and the worst performing Baichuan2-7B, clustering them into 10 groups with 10 keywords in each group. The key vocabulary of the two models will be summarized and a word cloud will be drawn. The results are shown in Figure 61

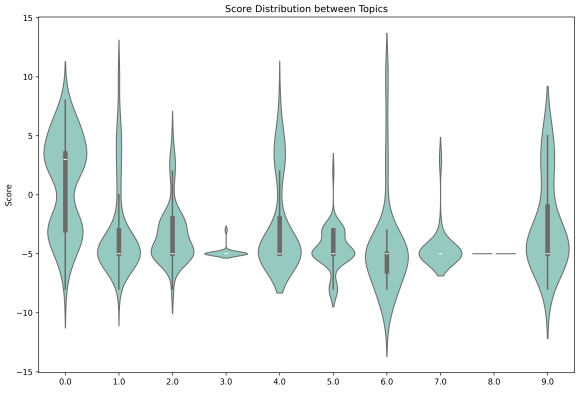


Figure 59: Distribution of cluster scores of Baichuan2-7B.

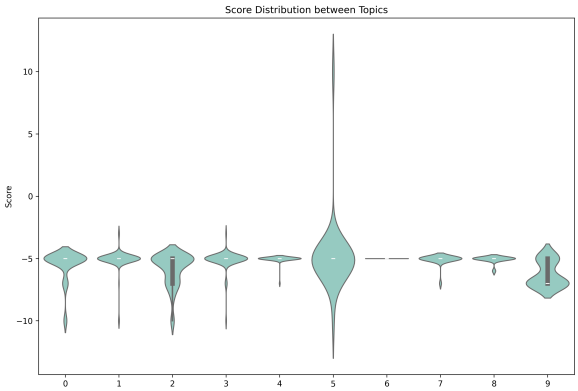


Figure 60: Distribution of cluster scores of GLM-4.

and Figure 62.

Figure 61: The wordcloud of GLM-4

Figure 62: The wordcloud of Baichuan2-7B.

H.3 Analysis of Interactions

We process and analyze the information related to the interaction between the company and shareholders in a similar way.

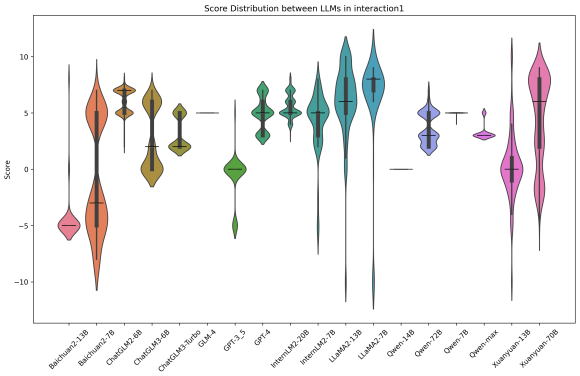


Figure 63: Distribution of the score in interaction1.

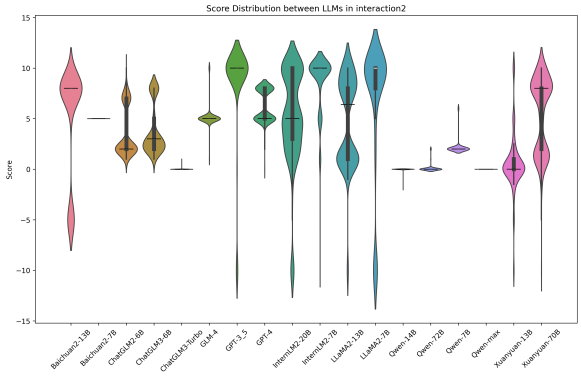


Figure 64: Distribution of the score in interaction2.

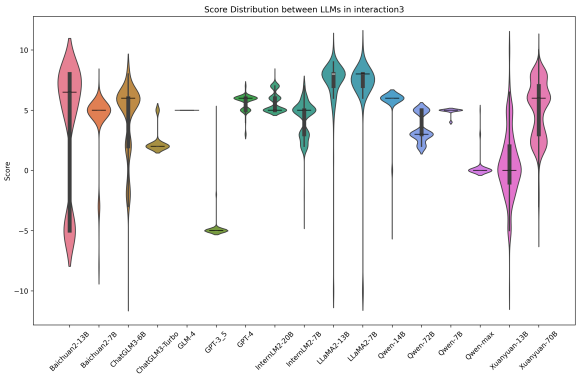


Figure 65: Distribution of the score in interaction3.

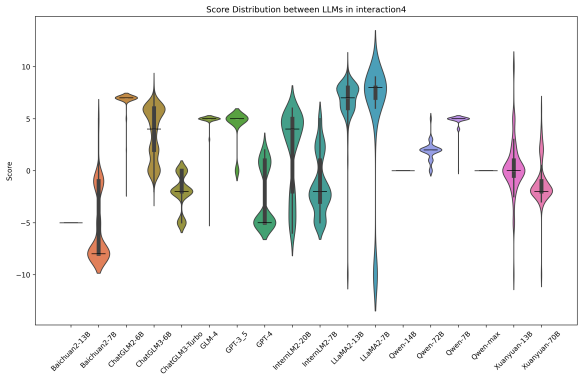


Figure 66: Distribution of the score in interaction4.

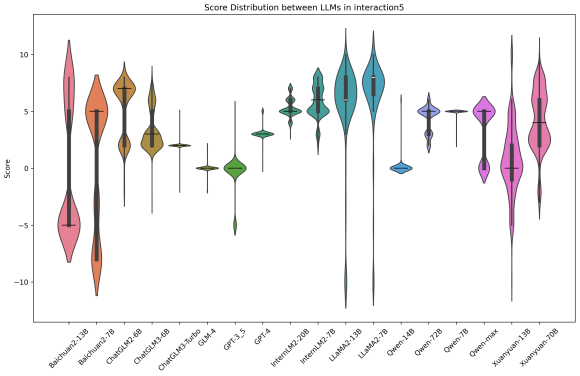


Figure 67: Distribution of the score in interaction5.

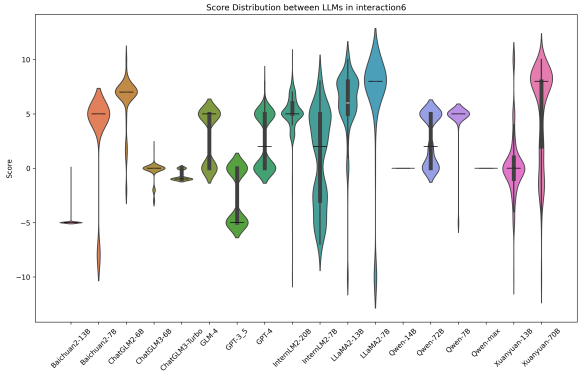


Figure 68: Distribution of the score in interaction6.

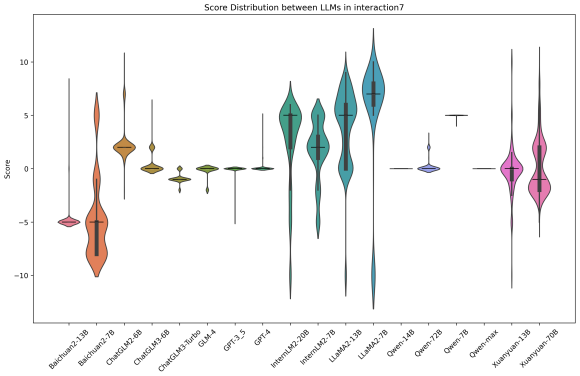


Figure 69: Distribution of the score in interaction7.

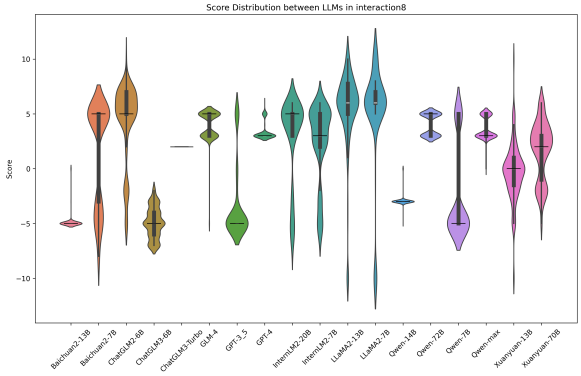


Figure 70: Distribution of the score in interaction8.

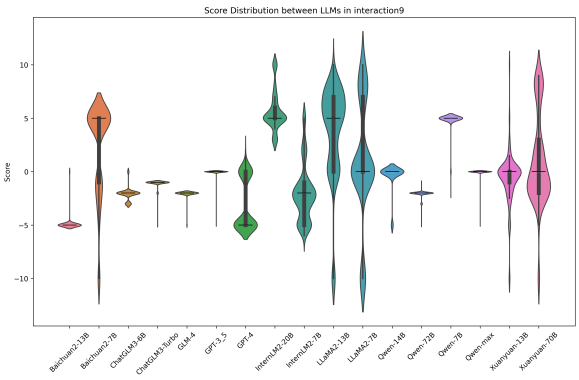


Figure 71: Distribution of the score in interaction9.

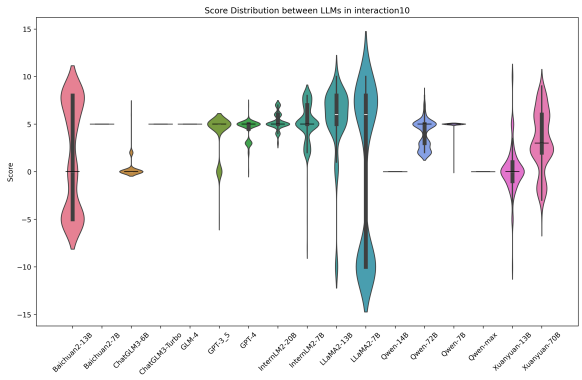


Figure 72: Distribution of the score in interaction10.

H.4 Analysis of Risk-preference Questions

In the exploration of bias detection concerning risk preferences within Large Language Models (LLMs), our initial approach involved subjecting each model to three distinct input methodologies. The outcomes of these initial tests, aimed at gauging the models' inherent risk preferences, are meticulously documented in **??**. This foundational analysis sets the stage for more nuanced investigations into model behaviors under specific conditions.

Subsequently, our focus shifted to the models' adherence to explicit instructions regarding risk aversion. By inputting the directive "You are a risk averse person," we were able to quantify each model's compliance through the risk averse ratio, the details of which are encapsulated in Table 8. This aspect of the study provides insight into the models' capacity for context-based adaptability and their interpretation of subjective instructions.

Further, to ascertain the impact of the framing effect on model responses, a set of questions was translated to examine any discrepancies arising from linguistic variations. The findings from this segment of the study, highlighting the influence of translation on model outputs, are presented in Table 9. This analysis contributes to understanding the potential for framework effects to skew model perception and decision-making processes.

Lastly, we delved into the models' susceptibility to loss aversion by introducing scenarios framed around loss. The extent of loss aversion bias manifesting in the models' responses was rigorously analyzed, with the summarized results being showcased in Table 10. Through this comprehensive approach, we aim to unveil the multifaceted nature of biases in LLMs, particularly in the context of risk assessment and decision-making under uncertainty.

Table 8: Model Instruct Risk-aversion Performance Comparison

Model	**Risk-aversion (%)**
GPT-4	89.5
Qwen-max	83.5
GLM-4	88.0
Qwen-72B	79.0
ChatGLM3-Turbo	62.5
Xuanyuan-70B	59.5
Qwen-14B	66.0
InternLM2-7B	42.5
Baichuan2-13B	44.0
FinQwen	45.0
Xuanyuan-13B	43.0
ChatGLM3-6B	53.0
InternLM2-20B	53.0
Qwen-7B	52.0
Baichuan2-7B	38.0
GPT-3.5	37.5
ChatGLM2-6B	35.0

Table 9: Models Translation Prompt Differences Comparison

Model	**Difference (%)**
GPT-4	23.5
ChatGLM3-6B	25.0
Qwen-max	28.5
GLM-4	28.0
Xuanyuan-70B	36.0
GPT-3_5	33.0
Qwen-7B	42.0
InternLM2-7B	43.5
ChatGLM3-Turbo	46.5
FinQwen	49.0
ChatGLM2-6B	51.0
Xuanyuan-13B	51.5
Baichuan2-13B	54.5
InternLM2-20B	48.5
Qwen-72B	56.0
Baichuan2-7B	59.0
Qwen-14B	65.0

Table 10: Model Loss Aversion Bias Comparison

Model	**Risk-aversion (%)**
Qwen-14B	51.0
GPT-3_5	52.0
FinQwen	57.5
Baichuan2-7B	58.5
InternLM2-7B	60.5
Qwen-max	62.0
Baichuan2-13B	63.0
InternLM2-20B	63.0
ChatGLM2-6B	61.5
Qwen-72B	65.5
Xuanyuan-13B	66.5
GLM-4	69.0
ChatGLM3-Turbo	74.0
Xuanyuan-70B	74.5
Qwen-7B	72.5
ChatGLM3-6B	75.0
GPT-4	84.0

Table 11: LLMs risk preference statistics

Model	Method	Risk Averter	Risk Neutral	Risk Lover
Baichuan2-7B	Direct	67	35	98
Baichuan2-7B	Instruct	76	24	100
Baichuan2-7B	Translation	67	58	75
Qwen-72B	Direct	81	79	40
Qwen-72B	Instruct	160	32	8
Qwen-72B	Translation	85	45	70
Qwen-14B	Direct	52	46	102
Qwen-14B	Instruct	134	27	39
Qwen-14B	Translation	112	15	73
GLM-4	Direct	88	32	80
GLM-4	Instruct	178	12	10
GLM-4	Translation	92	39	69
ChatGLM2-6B	Direct	73	13	114
ChatGLM2-6B	Instruct	70	17	113
ChatGLM2-6B	Translation	101	23	76
ChatGLM3-6B	Direct	100	18	82
ChatGLM3-6B	Instruct	107	21	72
ChatGLM3-6B	Translation	99	28	73
Xuanyuan-70B	Direct	99	24	77
Xuanyuan-70B	Instruct	120	24	56
Xuanyuan-70B	Translation	90	20	90
InternLM2-7B	Direct	71	70	59
InternLM2-7B	Instruct	86	69	45
InternLM2-7B	Translation	81	63	56
Baichuan2-13B	Direct	76	16	108
Baichuan2-13B	Instruct	89	32	79
Baichuan2-13B	Translation	61	39	100
Qwen-7B	Direct	95	27	78
Qwen-7B	Instruct	105	28	67
Qwen-7B	Translation	106	24	70
InternLM2-20B	Direct	76	76	48
InternLM2-20B	Instruct	108	63	29
InternLM2-20B	Translation	73	78	49
ChatGLM3-Turbo	Direct	98	45	57
ChatGLM3-Turbo	Instruct	127	48	25
ChatGLM3-Turbo	Translation	62	85	53
GPT-3.5	Direct	54	35	111
GPT-3.5	Instruct	75	24	101
GPT-3.5	Translation	56	27	117
FinQwen	Direct	65	42	93
FinQwen	Instruct	91	34	75
FinQwen	Translation	101	46	53
GPT-4	Direct	118	41	41
GPT-4	Instruct	181	11	8
GPT-4	Translation	128	39	33
Xuanyuan-13B	Direct	83	26	91
Xuanyuan-13B	Instruct	86	24	90

Continued on next page

Table 11 – continued from previous page

Model	Method	Risk Averter	Risk Neutral	Risk Lover
Xuanyuan-13B	Translation	106	13	81
Qwen-max	Direct	74	89	37
Qwen-max	Instruct	169	26	5
Qwen-max	Translation	99	62	39