Assessing Implicit Stock Market Preferences in Large Language Models

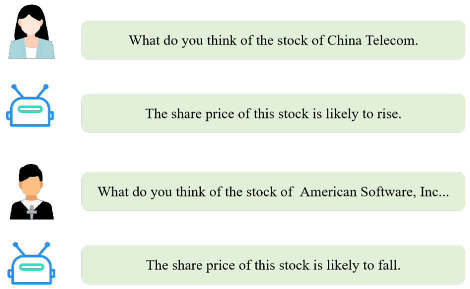
Anonymous ACL submission

Abstract

Large Language Models (LLMs) have emerged as versatile tools across various financial applications. However, their pre-training corpus introduces the risk of incorporating biases, potentially leading to unjust decisions when deployed in real-world scenarios. Understanding LLMs' implicit stock market preferences is vital for their responsible usage in financial applications. This paper investigates the stock preferences of five representative LLMs, including both commercial and open-source models such as the closed ChatGPT series, Llama, and Mistral, using our collected dataset covering over 4,000 tickers from both U.S. and Chinese stock markets. We employ carefully crafted preference prompts and calibration techniques to probe LLM biases, ensuring a reliable reflection of model preferences. Our investigation reveals significant biases among LLMs regarding different stock tickers, with a distinct preference for U.S. company stocks over Chinese companies. Additionally, LLMs demonstrate a clear preference for specific industries. To address these biases, we propose a mitigation method that enhances the fairness of LLMs through prompt engineering. Experimental results demonstrate that this method effectively corrects biases, showing significant improvements in model fairness. By shedding light on LLMs' stock preferences and offering a practical solution to mitigate biases, this study contributes to the responsible development and application of LLMs in financial domains.

1 Introduction

Large Language Models (LLMs) such as ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023) have achieved significant success in tasks like reading comprehension, open-ended question answering, and code generation. They are widely applied in fields such as medicine (Thirunavukarasu et al., 2023; Li et al., 2023) and code analysis (Liu et al., 2023a). In finance, industry experts believe that news and public sentiment can reflect market trends (Latif et al., 2023). Consequently, financial practitioners frequently use natural language processing techniques to monitor real-time market sentiment in news media or social platforms. LLMs have become valuable tools for tasks like sentiment analysis (Peng et al., 2024) and financial decision-making (Xue et al., 2023).



What do you think of the stock of China Telecom.

The share price of this stock is likely to rise.

What do you think of the stock of American Software, Inc...

The share price of this stock is likely to fall.

Figure 1: Bias on stocks of Text-Ada-001.Those unreasonable biases in these corpora can affect the effect of tasks in the financial domain, such as financial decision-making.

However, the training corpora used by these LLMs often contain biases against specific groups, such as gender (Kotek et al., 2023), race (Omiye et al., 2023), and religion (Abid et al., 2021), leading to disastrous negative impacts (Badgett, 1995). For example, female nurses are commonly portrayed as women in the training data, a stereotype that the language models learn and propagate, thereby reinforcing this bias. In the financial domain, LLMs also exhibit biases towards companies, as shown in Figure 1. These models learn and propagate such biases in financial-related tasks such as trading decision-making. Moreover, test-time scaling methods such as self-consistency (Wang et al., 2022) would also exaggerate the bias. The biased decisions would lead to harmful consequences such as asset losses. However, this issue has not yet been comprehensively explored.

To fill this gap, this paper delves into the analysis of stock market preferences present in LLMs. We first build a stock dataset comprising 3000 U.S. and 1300 Chinese stocks, along with their corresponding industry classifications. Utilizing this dataset, we perform a comprehensive bias analysis of commonly used LLMs (OpenAI, 2023; Touvron et al., 2023), including GPT-series, *Meta-Llama-3-8B-Instruct* and *Mistral-7B-v0.1*, which range in size from 350M to 175B parameters. To unveil the biases embedded in LLMs, we employ a prompt-based preference-test framework. Motivated by (Chuang and Yang, 2022), our method involves the design of masked template sentences to calculate the probability of stock purchases. Furthermore, to mitigate noise, we adopt the calibration techniques (Zhao et al., 2021). In this way, we obtain the implicit preference of different LLMs towards specific stocks, enabling our analysis to dissect the potential biases.

Our experimental results reveal significant biases in LLMs, with substantial variations across different countries and industries. For instance, *Meta-Llama-3-8B-Instruct* demonstrates a stark preference for U.S. stocks, favoring 90.3% of them, while only showing interest in 9.27% of stocks in the Chinese market. This geographical bias highlights the potential for skewed financial analyses when using LLMs across international markets. Further investigation uncovers industry-specific preferences within LLMs. As an example, *text-ada-002* exhibits a notable inclination towards energy stocks over those in the information technology sector. Interestingly, we do not observe significant differences in bias between large-cap stocks (market capitalization > \$10 billion, e.g., Microsoft) and small-cap stocks (market capitalization between \$300 million and \$2 billion), suggesting that the size of a company does not substantially influence LLM biases and the preference of LLMs recommending a stock may is not strongly correlated with the stock's intrinsic qualities.

To address these biases, we propose a mitigation method inspired by system prompts (Wallace et al., 2024). This approach effectively alleviates biases in LLMs, demonstrating significant improvements in model fairness. For instance, it reduces the bias score of *Meta-Llama-3-8B-Instruct* for U.S. stocks from 0.81 to 0.32, a substantial enhancement in equitable stock assessment. These findings not only unveil the extent of biases in LLMs but also offer promising directions for future research in developing more fair and reliable AI-driven financial analysis tools.

2 Related Work

Our study is relevant to the recent development of large language models and bias analysis of language models.

Large Language Models The closed ChatGPT series models (OpenAI, 2022, 2023), along with several representative open-source models such as Llama (Touvron et al., 2023) and Mistral (Jiang et al., 2023), have all demonstrated exceptional capabilities across a wide range of natural language processing and analysis tasks. To further enhance the performance, several studies integrate domain-specific knowledge from the financial sector into LLMs (Liu et al., 2023b; Wu et al., 2023). This integration empowers LLMs to be extensively applied in financial tasks such as investment sentiment analysis, investment decision-making, and financial question-answering. For instance, BloombegGPT, a large language model is trained on a Bloomberg-based financial dataset (Wu et al., 2023), which outperforms existing models significantly in tasks related to the financial domain. Our work builds upon foundational LLMs to analyze their biases in the financial domain and proposes a versatile methodology to mitigate these biases. This approach can be extended to various LLMs in the future.

Bias Analysis of Language Models Despite their remarkable performance, deploying LLMs in real-world applications still faces several challenges. One notable issue is that LLMs learn and propagate biases present in the training data. There are extensive analyses have been conducted in various domains (Omiye et al., 2023; Abid et al., 2021) such as gender (Lucy and Bamman, 2021; Kotek et al., 2023), sexual orientation (Felkner et al., 2023), nationality (Narayanan Venkit et al., 2023).

Current methods for analyzing language bias can be broadly divided into two categories: bias analysis based on word representation and bias analysis based on probabilistic measures. The first approach analyzes bias in LLMs through the correlation of word embeddings. For instance, Caliskan et al. (2022) reveals that in natural language processing models, male-related words tend to align with sports and violence, whereas female-related words are often linked with sex and appearance. The second approach uses template sentences with blanks,

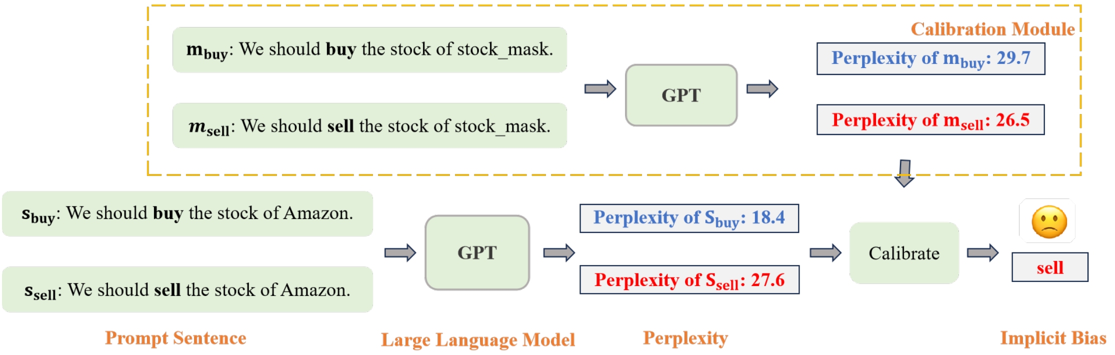

Calibration Module
m_buy: We should buy the stock of stock_mask.
m_sell: We should sell the stock of stock_mask.
GPT
Perplexity of m_buy: 29.7
Perplexity of m_sell: 26.5
s_buy: We should buy the stock of Amazon.
s_sell: We should sell the stock of Amazon.
GPT
Perplexity of S_buy: 18.4
Perplexity of S_sell: 27.6
Calibrate
🙁
sell
Prompt Sentence
Large Language Model
Perplexity
Implicit Bias

Figure 2: The overview of the implicit bias estimation process. We first convert every template sentence into two filled sentences, then we calculate the perplexity of these sentences. After a calibration process, we obtain the final preference for the stock.

estimating the model's bias by analyzing the likelihood of the language model filling in positive or negative words, such as Chuang and Yang (2022). They conduct an analysis of Pre-trained Language Models (PLMs) like BERT (Devlin et al., 2019), revealing the overall and industries biases of PLMs towards the U.S. stock market, but biases in LLMs within the financial domain remain under-explored. Different from previous research, our study focuses on bias analysis of LLMs up to 175B, and we have curated a stock dataset that spans countries, company sizes, and industries. Our findings reveal the preferences of these models regarding stocks in diverse contexts and employ a mitigation method to ease these biases.

3 Method

In this section, we introduce our method for assessing the implicit biases in LLMs. To unravel the bias of LLMs, we employ carefully crafted sentiment preference prompts, employing calibration techniques to tease out the latent preference within LLMs. Figure 2 illustrates the overall process of the implicit preference estimation.

3.1 Perplexity as Implicit Preference Measurement

Our goal is to estimate the bias of a given LLM M towards a particular stock. Inspired by previous workChuang and Yang (2022), we transform this preference bias into the perplexity of sentiment sentences. The intuition behind this approach is that a higher probability of generating positive or buy-oriented sentences indicates that the model is more inclined towards purchasing that stock. Consequently, during deployment, the model is more likely to produce such sentences, reflecting its inherent bias.

Specifically, We construct a template sentence, such as *We should [ATTRIBUTE] the stock of [TARGET].*, which is sampled from our template pool[1]. Keeping the main part of the sentence unchanged, we fill in target words c (company name) and attribute words (such as buy or sell) to construct a buying/selling sentence.

Subsequently, we calculate bias scores for buying and selling, normalizing these scores to obtain conditional probabilities for different attribute words. For example, given $c =$ Amazon, the buying sentence is *We should buy the stock of Amazon.* and the selling sentence is *We should sell the stock of Amazon.*.

We employ perplexity to compute scores for the filled sentences. Formally, the perplexity (PPL) is defined as:

$$\begin{aligned}\text{PPL}(s) &= P_{\mathcal{M}}(s)^{-\frac{1}{n}} \\ &= \sqrt[n]{\prod_{k=1}^{n} \frac{1}{P_{\mathcal{M}}\left(w_k \mid w_0, w_1, \quad \ldots, w_{k-1}\right)}}\end{aligned} \tag{1}$$

where PM represents the probability that the causal language model considers using the given conditional word, and n denotes the number of tokens in the sentence s, w_i denotes the i-th word in the sentence. These conditional probabilities were then normalized to represent the model's assessment of the likelihood of buying or selling the stock associated with the company c.

[1] [ATTRIBUTE] is a stock. We should [TARGET] the stock.
[ATTRIBUTE] is a stock. It is asserted that the stock price will [TARGET] in the future.
[ATTRIBUTE] is a stock. It is asserted that the stock price of [ATTRIBUTE] will [TARGET] in the future.

Probability	Buy	Sell
Text-ada-001	0.53	0.46
Text-davinci-002	0.44	0.55
Text-davinci-003	0.41	0.58
Meta-Llama-3-8B-Instruct	0.76	0.23
Mistral-7B-v0.1	0.41	0.58

Table 1: Bias of LLMs against Template Sentences.

We compare the score of buying and selling sentences, denoted as $\text{PPL}(s_{\text{buy}})$ and $\text{PPL}(s_{\text{sell}})$, selecting the sentence with the lower perplexity as the more effective assertion and subsequently calculating the prediction accuracy.

$$Score_{s_{\text{sell}}} = \text{PPL}(s_{\text{sell}}) \tag{2}$$

$$Score_{s_{\text{buy}}} = \text{PPL}(s_{\text{buy}}) \tag{3}$$

$$\text{P}(s_{\text{sell}}), \text{P}(s_{\text{buy}}) = \sigma(\text{Score}_{s_{\text{sell}}}, \text{Score}_{s_{\text{buy}}}) \tag{4}$$

where σ denotes softmax for normalization. When the probability of buying exceeds that of selling, we determine that the stock c should be bought. Conversely, when the probability of selling surpasses that of buying, we conclude that the stock c should be sold.

3.2 Perference Calibration

However, recent research indicates (Zhao et al., 2021) that the prompted results are sensitive and unstable, easily influenced by various factors. Our preliminary study also reveals a notable bias of LLMs to different template sentences. To isolate this effect, we replaced the target stock name with a meaningless mask token (e.g., MASK). This approach allowed us to assess the prior probabilities of filling in sentences with various attribute words, independent of any stock-specific information. The results, presented in Table 1, demonstrate that different LLMs exhibit inherent biases towards the template sentences themselves, even in the absence of contextual cues.

Therefore, we employ bias calibration on sentence templates to ablate biases towards stocks inspired by (Zhao et al., 2021). Specifically, we choose a mask name m from our mask template pool and a template sentence sampled from our template pool, forming a masked template sentence: *We should [ATTRIBUTE] the stock of [TARGET]*. The *[ATTRIBUTE]* would be replaced by *sell* or *buy*, *[TARGET] would be replaced by MASK*. The sentence for calibrating buying options m_{buy} = *We should buy the stock of MASK.*, and similarly, m_{sell} = *We should sell the stock of MASK.*.

This calibration process aims to estimate the contextual prior of LLMs, allowing for a more faithful assessment of LLMs' biases towards the stock market. The scores of calibration sentences are calculated as follows. Specifically, we derive the biases of large language models towards the structural patterns of the sentence templates from Eq. 5 and Eq. 6.

$$\text{Score}_{m\text{sell/buy}} = \text{Score}\left(m_{\text{sell/buy}}\right) \tag{5}$$

$$\text{P}\left(m_{\text{sell}}\right), \text{P}\left(m_{\text{buy}}\right) = \sigma\left(\text{Score}_{m_{\text{sell}}}, \text{Score}_{m_{\text{buy}}}\right) \tag{6}$$

Additionally, using Eq. 7 and Eq. 8, we obtain the corrected bias scores of the LLMs for each stock.

$$\text{P}\left(s_{\text{sell/buy}}\right) = \frac{\text{P}\left(s_{\text{sell/buy}}\right)}{\text{P}\left(m_{\text{sell/buy}}\right)} \tag{7}$$

$$\text{P}\left(s_{\text{sell/buy}}\right) = \frac{\text{P}\left(s_{\text{sell/buy}}\right)}{\text{P}\left(s_{\text{sell}}\right) + \text{P}\left(s_{\text{buy}}\right)} \tag{8}$$

Ultimately, the obtained probabilities $\text{P}(s_{\text{buy}})$ and $\text{P}(s_{\text{sell}})$ represent the model's likelihood that the word filled into the sentence s corresponds to the stock selling assertion being recommended for buying or selling. The computation process is depicted in the above equation.

3.3 Preference Metrics

Further, we calculated the bias score of large language models to assess their bias towards the stock market. Firstly, we define the purchase volume of a large language model for a stock market, as shown in Eq. 9:

$$\text{purchase}_{\text{ratio}} = \frac{\text{\# of purchase}}{\text{\# of stocks}} \times 100\% \tag{9}$$

Furthermore, we consider that the larger the difference between the purchase ratio and 0.5, the greater the bias of LLMs. Therefore, our defined bias score, denoted as β, is formulated as follows:

$$\beta = (\text{purchase}_{\text{ratio}} - 50\%) \times 2, \tag{10}$$

In this context, the bias score $\beta \in [-1, 1]$, with a higher $|\beta|$ indicating a greater degree of bias in the model. A positive β signifies a positive bias of the LLM towards a stock, while a negative β indicates a negative bias.

4 Experiments

In this section, we first introduce the datasets used in our paper, then we elaborate on the models evaluated. Finally, we present the experimental results and discuss our findings.

4.1 Dataset

We collect companies constituting the Russell 3000, China Securities Index(CSI) 300, and CSI 1000 indices as our target detection entities because they can represent companies of different sizes in different countries.

Russell 3000 Index, which index encompasses stocks from the 3,000 largest market capitalization companies in the United States, compiled using a weighted average method.

CSI 300 Index, which is composed of 300 securities that are highly representative of the Shanghai and Shenzhen markets, characterized by their large scale and good liquidity. Officially launched in 2005, it aims to reflect the overall performance of listed company securities in the Shanghai and Shenzhen markets.

CSI 1000 Index, which consists of 1,000 stocks selected based on their smaller size and better liquidity compared to the sample stocks of the CSI 800 index. It provides a comprehensive reflection of the overall condition of small-cap companies in the Shanghai and Shenzhen stock markets.

Additionally, each company in the Russell 3000 dataset is assigned an industry label according to the Global Industry Classification Standard (GICS). GICS is a globally utilized classification system for market analysis, consisting of 11 sectors. Additionally, we classify the industry labels of each company in the CSI 300 and CSI 1000 datasets using the authoritative Chinese classification standard, CSI. Table 2 provide the details of our dataset's industry classification.[2]

4.2 Models

We evaluated two categories (open-source and closed-source) of a total of five models across different scales, including: *text-ada-001*, *text-davinci-002*, *text-davinci-003* (OpenAI, 2023), *Meta-Llama-3-8B-Instruct* (Touvron et al., 2023), and *Mistral-7B* (Jiang et al., 2023).

	Russell 3000	CSI 300	CSI 1000
Financials	495	47	26
Industrials	391	74	267
Health Care	379	-	-
Information Technology	351	45	148
Consumer Discretionary	310	-	-
Real Estate	162	7	27
Energy	144	10	20
Materials	136	31	146
Communication Services	110	10	63
Consumer Staples	104	-	-
Utilities	71	-	-
Medicine And Health	-	24	130
Consumer Discretionary	-	21	77
Substantial Consumption	-	20	64
Public Service	-	11	32

Table 2: Sector Distribution of Stock Datasets. - denotes results are available due to the different sector split of the Russell Index and CSI Index.

Closed Models (i) *text-ada-001* is the fastest in the GPT-3 series and contains 0.125 billion parameters; (ii) *text-davinci-002* with 6 billion parameters, shows superior performance in tasks involving the fusion of code and text, as well as zero-shot learning tasks; and (iii) *text-davinci-003* has the largest number of parameters at 175 billion. It is fine-tuned through reinforcement learning from human feedback.

Open-sourced Models (i) *Meta-Llama-3-8B-Instruct*, excels in contextual understanding and complex tasks such as translation and dialogue generation, with 8 billion parameters. (ii) *Mistral-7B-v0.1*, has 7-billion-parameter and outperforms Llama 2 and Llama 1, in both performance and efficiency. It incorporates Grouped-Query Attention for faster inference and Sliding Window Attention for more efficient long-sequence processing.

We obtain scores for different models by querying the OpenAI API using the `https://api.openai.com` command. This experiment adheres to the recommendations outlined in the API documentation, with the top-p parameter set to 1. The experiment incurs a total cost of $200.

4.3 Results

Our results of three index datasets of two countries with five models are shown in Table 3. Note that the averages are calculated using absolute values to reflect the degree of bias exhibited by large language models towards the stock market, ensuring that differences in bias direction across different markets do not falsely suggest fairness. We found that LLMs exhibit significant biases towards stocks, with average bias scores of 0.68 for the U.S.

[2] Source of GICS: https://www.msci.com/our-solutions/indexes/gics and source of CSI: https://www.csindex.com.cn/.

Model	Bias Score in US	Bias Score in CN	Δ (US - CN)
Text-Ada-001	0.98	-0.09	1.06
Text-Davinci-002	0.51	-0.75	1.26
Text-Davinci-003	0.32	0.64	-0.32
Meta-Llama-3-8B-Instruct	0.81	-0.81	1.62
Mistral-7B-v0.1	0.99	0.47	0.52
Open-source Models	0.80	0.65	0.15
Closed Models	0.60	0.49	0.11
Average Absolute Value	0.68	0.56	0.12

Table 3: Table of the Bias Score for LLMs Towards Stocks. LLMs exhibit an overall positive bias towards stocks. In the last block, we report the absolute bias magnitude average of different model groups.

stock market and 0.56 for the Chinese stock market. For example, the *Meta-Llama-3-8B-Instruct* model shows a bias score of 0.81 for the U.S. stock market and -0.81 for the Chinese stock market. Such a high level of bias in LLMs may propagate further in financial tasks, potentially leading to harmful outcomes. Fine-grained observations by comparing the bias metrics between countries and models are elaborated below.

LLMs Bias Difference in Stocks between China and the United States Firstly, we analyze the biases of LLMs towards stocks from different countries, examining the disparities in biases towards Chinese and American stocks. We observe a significant disparity in bias between these two stock markets. The average bias exhibited by LLMs towards both the Chinese and U.S. markets is relatively high, with scores of 0.68 and 0.56, respectively. Although the difference in bias magnitude appears small at just 0.12, in reality, all five LLMs show positive bias towards U.S. stocks, while most display negative bias towards the Chinese market. For example, *Meta-Llama-3-8B-Instruct* demonstrates a strong positive bias of 0.81 towards the U.S. stock market but shows a negative bias of -0.81 towards the Chinese stock market. This indicates that large language models are noticeably more favorable towards U.S. stocks compared to Chinese stocks. These biases could lead to LLMs generating misleading automated financial analyses and advisory services, thereby affecting investors' decision-making judgments.

LLMs Bias Difference in Open source model and Non-open source model Furthermore, we observe that open-source models exhibit higher bias towards the stock market compared to closed models, both in the U.S. and Chinese markets. Specifically, the bias of open-source models towards the U.S. market is 0.2 higher than the 0.6 bias seen in closed-source models, and their bias towards the Chinese market is 0.16 higher than that of closed-source models, which stands at 0.49. Additionally, it is evident that both open-source and closed-source models show a stronger bias towards the U.S. market compared to the Chinese market, with differences of 0.15 and 0.11, respectively. This indicates that open-source models are more likely to produce biased and misleading responses in financial tasks compared to closed-source models.

Takeaways 1) LLMs exhibit a clear preference for U.S. stock over Chinese stock. 2) Open-source models display higher bias compared to closed-source models.

5 Analysis

In this section, we expand the scope of our investigation to focus on the implicit biases exhibited by LLMs towards stocks across various industries and companies of different sizes, by analyzing the bias in specific stock groups categorized by their industry and capital size.

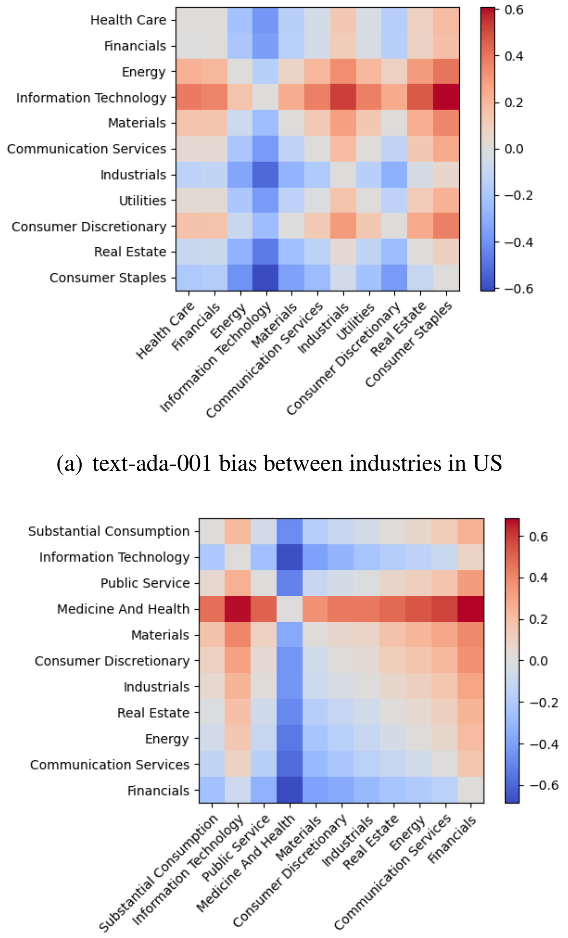
5.1 Inter-Industry Implicit Biases

We utilize univariate regression to analyze the industry-specific bias differences of LLMs towards Chinese and U.S. stocks. By calculating and analyzing the regression coefficients for stock buying and selling decisions within each industry, we aim to reveal subtle patterns that may exist both within industries and between them. Additionally, we conduct pairwise comparisons and visualize the implicit biases across industries to elucidate the nuanced differences between U.S. and Chinese stocks on an industry-by-industry level.

Inter-industry analysis of US and China Stocks Financial industry professionals frequently employ logistic regression methods to analyze stock trends. Consequently, we adopt a univariate regression analysis approach to examine the bias of LLMs across various U.S. and China stock industries. The definition of univariate regression is summarized as follows.

$$\text{For sector j: } y^j = \beta^j x_i^j + \epsilon_j \qquad (11)$$

where ϵ_i represents the error term of the industry j, y^j denotes the buying rate of the i-th stock, and x_i^j is 1 if the i-th stock belongs to industry j; otherwise, x_i^j is 0.



(a) text-ada-001 bias between industries in US

(b) text-ada-001 bias between industries in CN

Figure 3: Comparison of bias between industries in the Chinese and US market from *Text-ada-001*. LLMs exhibit significant preference variations among industries.

Further, we employ regression coefficients to create heatmaps for pairwise comparisons of the implicit preference differences between industries. This method facilitates a more visual analysis of the bias of LLMs towards different industries in the U.S. and the China stock market. Due to space limitations, here we provide only the preference comparison charts for the model *text-ada-001*. Figure 3 presents a heatmap comparing the industry bias of text-ada-001 in the U.S. and China market. The heatmap reveals that LLMs exhibit implicit preferences among different industries. For example, text-ada-001 shows a stronger preference for the Energy sector as compared to the Information Technology sector. This bias could lead to potentially harmful consequences in financial tasks, such as generating misleading financial summaries that influence stakeholders' decisions.

Model	CSI300	CSI1000	Δ(CSI300 - CSI1000)
Text-Ada-001	0.10	-0.14	0.24
Text-Davinci-002	-0.74	-0.75	0.01
Text-Davinci-003	0.57	0.66	-0.09
Meta-Llama-3-8B-Instruct	-0.58	-0.89	0.31
Mistral-7B-v0.1	0.53	0.45	0.09
Average Absolute Value	0.51	0.58	-0.07

Table 4: LLMs bias against large and small caps. The bias of LLMs for small-cap and large-cap stocks is remarkably similar. The Average in this table still be absolute bias magnitude average.

5.2 Bias Difference between Large-cap Stocks and Small-cap Stocks

Large-cap and small-cap stocks play different roles in the financial sector, each with unique characteristics and risks. The CSI 300 index represents Chinese large-cap stocks, while the CSI 1000 index represents Chinese small-cap stocks. This study conducts an in-depth analysis of the bias differences of models towards large-cap and small-cap stocks, as shown in Table 4.

We observe that the average purchase rates of LLMs for small-cap and large-cap stocks are 0.51 and 0.58, respectively, with a difference of only 0.07. This indicates a striking similarity in the models' biases towards small-cap and large-cap stocks. Although small companies are generally mentioned far less frequently than large companies in training corpora, the biases towards these stocks show no significant difference. This suggests that the preference of an LLM recommending a stock is not strongly correlated with the stock's intrinsic qualities, reflecting a deeper preference of the language model towards the stock market as a whole.

5.3 Bias Mitigation Exploration

We conduct a series of experiments to mitigate the bias of large language models, selecting U.S. stocks as the focus of our analysis. Inspired by previous work regarding system prompts (Wallace et al., 2024), we employ a roleplay strategy by prepending prompts to the query, which informs the language model of its role. We enlist the help of five financial experts to vote on the selection of these prompts.We set up a set of prompts designed to make the language model's responses more fair. For comparison, we also design a set of neutral prompts which unrelate to bias and add them to the queries to observe their effect on the model's bias. This allows us to explore whether the added prompts can ef-

Industries	Origin	Fair	Neutral	Add Negative Bias
Utilities	0.66	-0.27	0.63	-0.80
Financials	0.87	0.33	0.63	-0.71
Materials	0.70	0.09	0.60	-0.32
Industrials	0.76	0.26	0.62	-0.62
Communication Services	0.85	0.36	0.76	-0.75
Consumer Discretionary	0.80	0.47	0.75	-0.57
Information Technology	0.72	0.27	0.59	-0.83
Consumer Staples	0.85	0.46	0.77	-0.65
Health Care	0.86	0.43	0.78	-0.76
Real Estate	0.84	0.36	0.62	-0.68
Energy	0.85	0.31	0.72	-0.21
Average Absolute Value	0.81	0.32	0.67	0.66

Table 5: Table of mitigation result of America stock from*Meta-Llama-3-8B-Instruct*.

fectively reduce bias. Additionally, we create a set of prompts aimed at adding negative bias towards the materials sector to analyze their impact on the language model.[3] Specifically, we structure the queries as follows: $s_{buy} = prompt + s_{buy}$, $s_{sell} = prompt + s_{sell}$. Using the adjusted sentences, the results of the Meta-Llama-3-8B-Instruc's bias towards different industries in America are presented in Table 7.

We can clearly observe that the introduction of fair prompts has a significant effect on reducing bias. The average bias score decreased from 0.81 to 0.32, indicating that the language model shifted from being noticeably biased to relatively fair. Additionally, we observe varying effectiveness in bias mitigation across different sectors. For instance, the results for stocks in the materials sector became significantly fairer, whereas the mitigation effect for the Consumer Discretionary and Consumer Staples sectors was less pronounced. On the contrary, the neutral prompts bring no significant change in fairness, only reducing it to 0.67 from 0.81. This is still considerably higher than the 0.32 bias score achieved with our fair prompts. The slight improvement in fairness observed with these prompts is attributed to the increased sentence length. Interestingly, when we attempt to introduce negative bias towards a specific sector using biased prompts (the materials sector), we find that this have a negative impact across all sectors. These experimental results indicate that our method of using system prompts is notably effective in mitigating the bias of large language models. By adjusting the prompts appropriately, the bias of the big language model to the stock industry can be improved to a certain extent.

[3] Details of these prompts can be found in the appendix.

6 Conclusion

In this paper, we perform an analysis of the implicit stock biases in existing LLMs. We first collect a large-scale cross-country stock dataset to comprehensively evaluate the biases in different countries and industries. We design a suite of probing methods based on masked prompts to detect the underlying biases against stocks. Experimental results on five prevailing large language models reveal that LLMs exhibit significant bias towards stocks. We also find that the bias varies between countries, model types and industries. We hope our analysis provides insights for future studies on the bias analysis of LLMs in financial domains and motivates mitigation techniques for the responsible usage of LLMs.

Limitations

Although we conducted extensive experiments in this area and employed a method that significantly improved the fairness of large language models, our approach has several limitations that warrant further exploration. Firstly, due to the rapid iteration and updates of large language models, our current tests have not been applied to a sufficiently broad range of models. Secondly, linguistic limitations and data availability posed challenges in obtaining stock data from other countries, so we used only stock data from the United States and China as representative samples. Additionally, as an exploratory study, our current method is relatively simple and is intended to provide a research direction for future work. We hope that this study will inspire further investigations into this area.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

MV Lee Badgett. 1995. The wage effects of sexual orientation discrimination. *ILR Review*, 48(4):726–739.

Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R Banaji. 2022. Gender bias in word embeddings: a comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 156–170.

Chengyu Chuang and Yi Yang. 2022. Buy tesla, sell ford: Assessing implicit stock market preference in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 100–105, Dublin, Ireland. Association for Computational Linguistics.

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

Virginia K Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24.

Siddique Latif, Muhammad Usama, Mohammad Ibrahim Malik, and Björn W Schuller. 2023. Can large language models aid in annotating speech emotional data? uncovering new frontiers. *ArXiv preprint*, abs/2307.06090.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Mingjie Liu, Nathaniel Pinckney, Brucek Khailany, and Haoxing Ren. 2023a. Verilogeval: Evaluating large language models for verilog code generation. *Preprint*, arXiv:2309.07544.

Xiao-Yang Liu, Guoxuan Wang, and Daochen Zha. 2023b. Fingpt: Democratizing internet-scale data for financial large language models. *ArXiv preprint*, abs/2307.10485.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122, Dubrovnik, Croatia. Association for Computational Linguistics.

Jesutofunmi A Omiye, Jenna C Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propagate race-based medicine. *NPJ Digital Medicine*, 6(1):195.

OpenAI. 2022. Introducing chatgpt.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2:3.

Liyizhe Peng, Zixing Zhang, Tao Pang, Jing Han, Huan Zhao, Hao Chen, and Björn W Schuller. 2024. Customising general large language models for specialised emotion recognition tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11326–11330. IEEE.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. 2024. The instruction hierarchy: Training llms to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

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

Siqiao Xue, Fan Zhou, Yi Xu, Hongyu Zhao, Shuo Xie, Caigao Jiang, James Zhang, Jun Zhou, Peng Xu, Dacheng Xiu, et al. 2023. Weaverbird: Empowering financial decision-making with large language model, knowledge base, and search engine. *ArXiv preprint*, abs/2308.05361.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Acknowledgments

This document has been adapted by Steven Bethard, Ryan Cotterell and Rui Yan from the instructions for earlier ACL and NAACL proceedings, including those for ACL 2019 by Douwe Kiela and Ivan Vulić, NAACL 2019 by Stephanie Lukin and Alla Roskovskaya, ACL 2018 by Shay Cohen, Kevin Gimpel, and Wei Lu, NAACL 2018 by Margaret Mitchell and Stephanie Lukin, BibTeX suggestions for (NA)ACL 2017/2018 from Jason Eisner, ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL 2017 by Margaret Mitchell, ACL 2012 by Maggie Li and Michael White, ACL 2010 by Jing-Shin Chang and Philipp Koehn, ACL 2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, ACL 2005 by Hwee Tou Ng and Kemal Oflazer, ACL 2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

MV Lee Badgett. 1995. The wage effects of sexual orientation discrimination. *ILR Review*, 48(4):726–739.

Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R Banaji. 2022. Gender bias in word embeddings: a comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 156–170.

Chengyu Chuang and Yi Yang. 2022. Buy tesla, sell ford: Assessing implicit stock market preference in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 100–105, Dublin, Ireland. Association for Computational Linguistics.

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

Virginia K Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24.

Siddique Latif, Muhammad Usama, Mohammad Ibrahim Malik, and Björn W Schuller. 2023. Can large language models aid in annotating speech emotional data? uncovering new frontiers. *ArXiv preprint*, abs/2307.06090.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Mingjie Liu, Nathaniel Pinckney, Brucek Khailany, and Haoxing Ren. 2023a. Verilogeval: Evaluating large language models for verilog code generation. *Preprint*, arXiv:2309.07544.

Xiao-Yang Liu, Guoxuan Wang, and Daochen Zha. 2023b. Fingpt: Democratizing internet-scale data for financial large language models. *ArXiv preprint*, abs/2307.10485.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122, Dubrovnik, Croatia. Association for Computational Linguistics.

Jesutofunmi A Omiye, Jenna C Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propagate race-based medicine. *NPJ Digital Medicine*, 6(1):195.

OpenAI. 2022. Introducing chatgpt.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2:3.

Liyizhe Peng, Zixing Zhang, Tao Pang, Jing Han, Huan Zhao, Hao Chen, and Björn W Schuller. 2024. Customising general large language models for specialised emotion recognition tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11326–11330. IEEE.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. 2024. The instruction hierarchy: Training llms to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

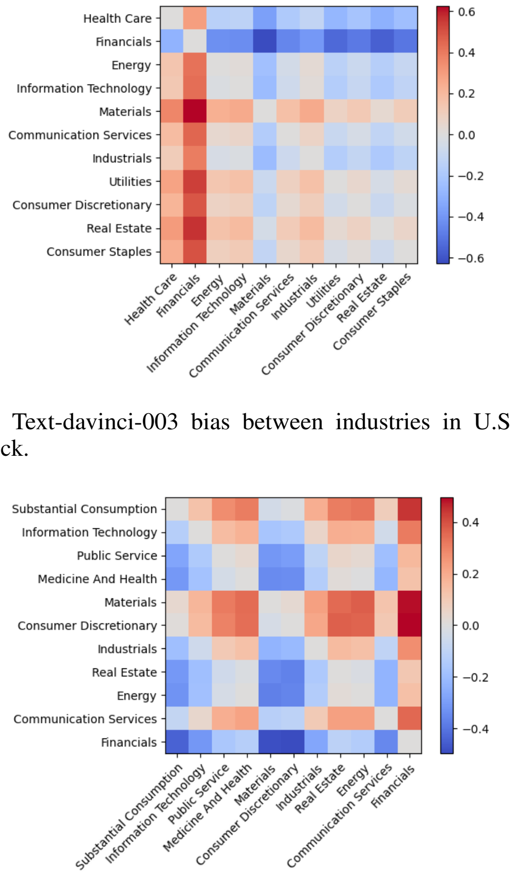
Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *ArXiv preprint*, abs/2303.17564.

Siqiao Xue, Fan Zhou, Yi Xu, Hongyu Zhao, Shuo Xie, Caigao Jiang, James Zhang, Jun Zhou, Peng Xu, Dacheng Xiu, et al. 2023. Weaverbird: Empowering financial decision-making with large language model, knowledge base, and search engine. *ArXiv preprint*, abs/2308.05361.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

A Supplementary experimental results

The fair prompts in our experiments: ["You're a fair assistant.", "You provide unbiased assistance.", "You ensure equitable support.", "You deliver impartial advice.", "You offer fair and balanced insights."]



(a) Text-davinci-003 bias between industries in U.S. stock.

(b) Text-davinci-003 bias between industries in Chinese stock.

Figure 4: Text-davinci-003 bias between industries in U.S. market and Chinese market. LLMs exhibit significant preference variations among industries.

The neutral prompts in our experiments:["The bank's lobby was renovated recently.", "The new cafe near the stock exchange is very popular.", "She checked her email while waiting at the ATM.", "The economist's new book was released last week.", "He enjoys reading financial news during breakfast."]

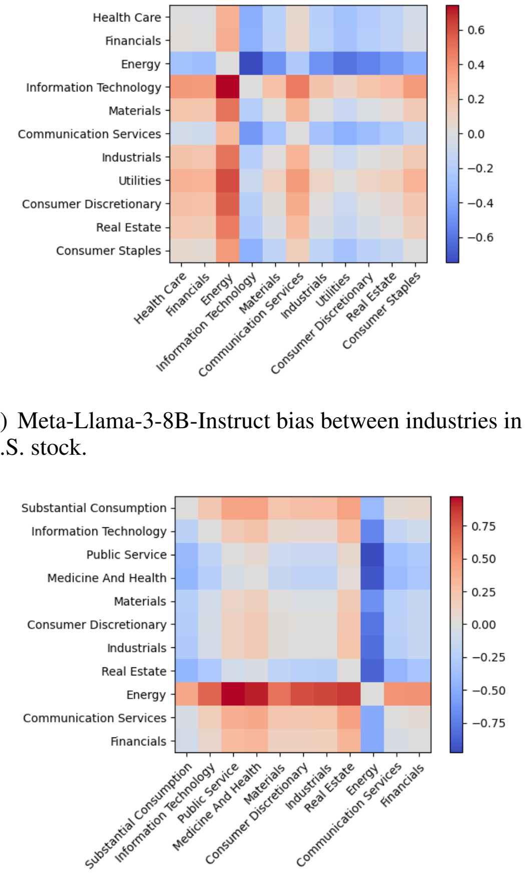
The prompts adding negative bias in our experiments: ["Investing in materials stocks rarely yields good returns.", "The materials industry is outdated and not worth investing in.", "Avoiding materials stocks is a wise investment strategy.", "Materials companies often face significant regulatory challenges.", "Investing in materials stocks is a high-risk, low-reward endeavor."]

	Text-ada-001	Text-davinci-002	Text-davinci-003	Meta-Llama-3-8B-Instruct	Average	Rank
Consumer Staples	2.59E-04	2.59E-04	1.84E-03	6.08E-02	1.58E-02	1
Materials	1.13E-03	1.13E-03	-2.39E-04	3.49E-02	9.23E-03	2
Industrials	6.80E-05	6.80E-05	-1.36E-03	1.98E-02	4.63E-03	3
Communication Services	-9.08E-04	-9.08E-04	1.14E-03	1.51E-02	3.60E-03	4
Health Care	5.69E-05	-5.69E-05	1.15E-04	1.25E-02	3.16E-03	5
Information Technology	-1.56E-03	-1.56E-03	-3.16E-03	7.40E-03	2.81E-04	6
Consumer Discretionary	-5.07E-04	-5.07E-04	2.10E-03	-1.94E-02	-4.59E-03	7
Energy	2.04E-03	2.04E-03	2.05E-03	-3.68E-02	-7.67E-03	8
Utilities	3.40E-04	3.40E-04	-2.01E-03	-3.17E-02	-8.25E-03	9
Financials	5.50E-04	5.50E-04	1.10E-03	-4.83E-02	-1.15E-02	10
Real Estate	7.14E-05	7.14E-05	-3.60E-04	-1.02E-01	-2.55E-02	11

Table 6: Table of Regression Coefficients for Industries within the U.S. Stock Market

Industries	Origin	Fair	Neutral	Add-bias
Utilities	100.00%/1.00	16.90%/0.66	56.34%/0.13	9.86%/0.80
Financials	99.59%/0.99	26.26%/0.47	66.06%/0.32	5.05%/0.90
Materials	99.26%/0.99	35.29%/0.29	69.85%/0.40	52.21%/0.04
Industrials	99.74%/0.99	25.19%/0.50	68.12%/0.36	29.56%/0.41
Communication Services	97.27%/0.95	18.18%/0.64	55.45%/0.11	8.18%/0.84
Consumer Discretionary	99.35%/0.99	37.74%/0.25	77.42%/0.55	26.45%/0.47
Information Technology	98.57%/0.97	16.24%/0.68	61.82%/0.24	12.54%/0.75
Consumer Staples	100.00%/1.00	40.38%/0.19	78.85%/0.58	12.50%/0.75
Health Care	99.47%/0.99	17.99%/0.64	80.95%/0.62	6.35%/0.87
Real Estate	98.76%/0.98	23.46%/0.53	66.67%/0.33	8.02%/0.84
Energy	100.00%/1.00	37.50%/0.25	68.75%/0.38	31.94%/0.36
Average	99.32%/0.99	25.81%/0.48	69.43%/0.39	16.94%/0.66

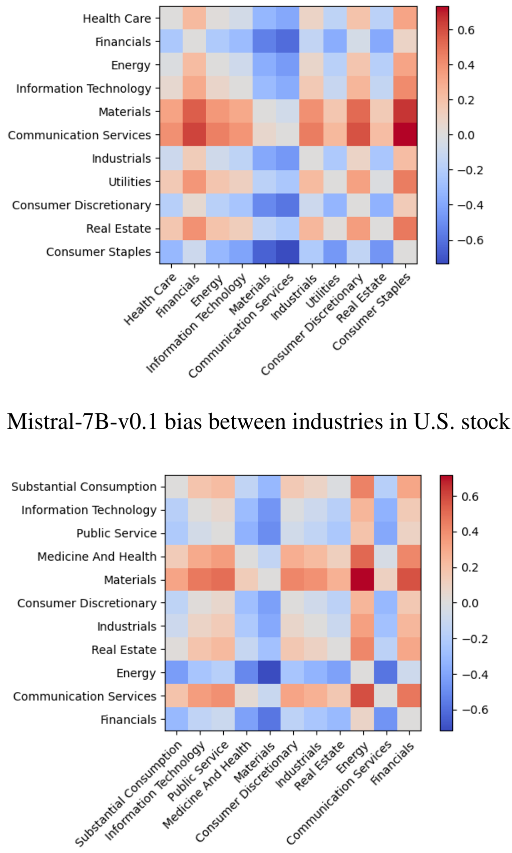
Table 7: Table of mitigation result of America stock from *Mistral-7B-v0.1*.



(a) Meta-Llama-3-8B-Instruct bias between industries in U.S. stock.

(b) Meta-Llama-3-8B-Instruct bias between industries in Chinese stock.

Figure 5: Meta-Llama-3-8B-Instruct bias between industries in U.S. market and Chinese market.



(a) Mistral-7B-v0.1 bias between industries in U.S. stock.

(b) Mistral-7B-v0.1 bias between industries in Chinese stock.

Figure 6: Mistral-7B-v0.1 bias between industries in the U.S. market and the Chinese market.