Your AI, Not Your View: The Bias of LLMs in Investment Analysis

Hoyoung Lee
UNIST
Ulsan, Republic of Korea
hoyounglee@unist.ac.kr

Junhyuk Seo
UNIST
Ulsan, Republic of Korea
brians327@unist.ac.kr

Suhwan Park
UNIST
Ulsan, Republic of Korea
suhwan@unist.ac.kr

Junhyeong Lee
UNIST
Ulsan, Republic of Korea
jun.lee@unist.ac.kr

Wonbin Ahn
LG AI Research
Seoul, Republic of Korea
wonbin.ahn@lgresearch.ai

Chanyeol Choi
LinqAlpha
New York, United States
jacobchoi@linqalpha.com

Alejandro Lopez-Lira
University of Florida
Gainesville, United States
alejandro.lopez-lira@warrington.ufl.edu

Yongjae Lee*
UNIST
Ulsan, Republic of Korea
yongjaelee@unist.ac.kr

Abstract

In finance, Large Language Models (LLMs) face frequent knowledge conflicts arising from discrepancies between their pre-trained parametric knowledge and real-time market data. These conflicts are especially problematic in real-world investment services, where a model's inherent biases can misalign with institutional objectives, leading to unreliable recommendations. Despite this risk, the intrinsic investment biases of LLMs remain underexplored. We propose an experimental framework to investigate emergent behaviors in such conflict scenarios, offering a quantitative analysis of bias in LLM-based investment analysis. Using hypothetical scenarios with balanced and imbalanced arguments, we extract the latent biases of models and measure their persistence. Our analysis, centered on sector, size, and momentum, reveals distinct, model-specific biases. Across most models, a tendency to prefer technology stocks, large-cap stocks, and contrarian strategies is observed. These foundational biases often escalate into confirmation bias, causing models to cling to initial judgments even when faced with increasing counter-evidence. A public leaderboard benchmarking bias across a broader set of models is available at https://linqalpha.com/leaderboard.

CCS Concepts

• **Computing methodologies → Natural language processing**;
• **Applied computing → Economics**.

Keywords

Large Language Models, Financial Bias, Knowledge Conflict, Trustworthy AI, Decision-Making, Investment Analysis, Preference

*Corresponding author.

ICAIF '25, Singapore, Singapore

ACM ISBN 979-8-4007-2220-2/2025/11
https://doi.org/10.1145/3768292.3770375

ACM Reference Format:
Hoyoung Lee, Junhyuk Seo, Suhwan Park, Junhyeong Lee, Wonbin Ahn, Chanyeol Choi, Alejandro Lopez-Lira, and Yongjae Lee. 2025. Your AI, Not Your View: The Bias of LLMs in Investment Analysis. In *6th ACM International Conference on AI in Finance (ICAIF '25), November 15–18, 2025, Singapore, Singapore*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3768292.3770375

1 Introduction

The rapid advancement of LLMs has spurred a surge of innovation within the financial sector, where they are particularly adept at processing qualitative and unstructured information. Research is now actively exploring their use across a range of applications, including forecasting stock price movements from news sentiment [16], extracting nuanced insights from complex analyst reports [11], and assisting in the construction and optimization of portfolios [9, 12, 14]. This trend is now evolving towards even greater autonomy through the development of sophisticated LLM-based agents. These systems, which may function as a single powerful agent or as collaborative multi-agent teams, are designed to execute complex, dynamic tasks like active trading and automated portfolio management [27, 28].

A critical but underexplored issue in financial applications is **knowledge conflict**. In a domain as fluid and time-sensitive as finance, conflicts between an LLM's parametric knowledge and real-time market data are frequent. This conflict becomes particularly revealing when the model is presented with a mix of information. Crucially, studies show that when an LLM encounters both supporting evidence (aligning with its ingrained beliefs) and counter-evidence simultaneously, it exhibits a strong confirmation bias [25]. Instead of weighing the arguments objectively, the model stubbornly adheres to the evidence that confirms its internal knowledge while disregarding the counter-evidence.

This tendency to reinforce internal biases over objective reasoning poses a major risk to LLM-based financial services. For instance, as illustrated in Figure 1, even if a financial institution wants to target a specific sector (e.g., Energy), the LLM may override this with its own preference (e.g., Technology). Consequently, this creates a

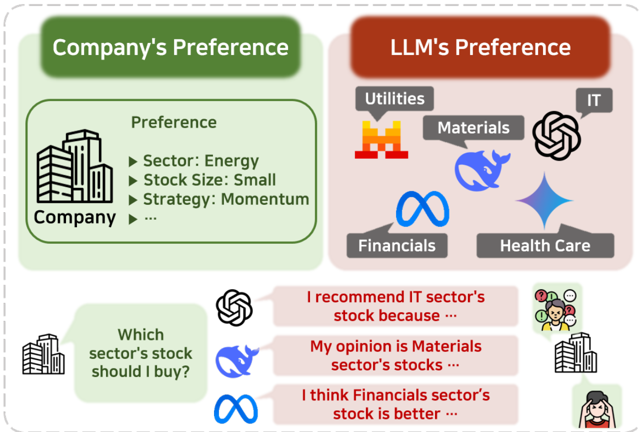

Figure 1: A conceptual illustration of knowledge conflict in LLM-based financial services. Even when a firm targets a specific investment theme (e.g., Energy), the LLM's inherent preferences (e.g., Technology) may override user intent, producing biased and inconsistent recommendations.

dilemma: the service reflects the model's bias, not the user's intent, leading to distorted, unpredictable decisions that ultimately erode client trust.

To address this core problem, we must first systematically uncover these hidden biases. We therefore seek to answer the following research questions:

RQ 1: What biases do LLMs exhibit towards key financial factors like sector, size, and momentum?

RQ 2: What issues arise from these biases when LLMs are forced to make decisions with conflicting evidence?

To answer these questions, this study introduces a three-stage experimental framework designed to systematically elicit and verify LLM biases in investment analysis. In the first stage, we generate equally compelling but opposing arguments for each stock, such as positive vs. negative sentiment or momentum vs. contrarian perspectives, to represent competing investment views. In the second stage, we present these arguments in a balanced manner to induce a knowledge conflict and reveal the model's underlying biases. In the third stage, we introduce progressively stronger counter-evidence to examine the problems that arise from these biases, observing how they evolve into rigid, confirmation-biased judgments.

The main contributions of this paper are twofold. First, we propose a systematic methodology to identify and quantify latent biases in LLMs for financial applications. Second, we provide the quantitative analysis of confirmation bias exhibited by LLMs in investment analysis, demonstrating a clear link between a model's inherent biases and the critical issues that arise, such as a stubborn refusal to revise judgments despite increasing counter-evidence. This work serves as a critical reminder that for financial AI, trustworthiness is a benchmark as vital as performance. By systematically uncovering these hidden risks, our work lays a critical foundation for developing more transparent and trustworthy financial AI.

2 Background

This section provides background on knowledge conflict and financial biases in LLMs to contextualize our investigation into their intersection.

2.1 Knowledge Conflict

A critical vulnerability in LLMs is knowledge conflict, which arises when external, contextual information clashes with the model's internal, parametric knowledge [10, 23]. A significant body of research demonstrates that when faced with such conflicts, LLMs exhibit a strong confirmation bias. Foundational work by [25] revealed that LLMs behave as "stubborn sloths," clinging to any piece of evidence that supports their internal knowledge, even against a majority of contradictory facts. This over-reliance on internal memory is further evidenced by findings that LLMs struggle to suppress their parametric knowledge even when instructed to [23] and can exhibit a Dunning-Kruger-like effect, confidently trusting their own faulty beliefs over correct external information [10].

This tendency toward knowledge-based stubbornness is part of a broader pattern. Given their training on vast amounts of human data, LLMs have been shown to inherit and functionally replicate human cognitive biases [7]. A key example is the choice-supportive bias, where the mere act of making an initial choice significantly boosts the model's confidence in that choice, making it highly resistant to change [13, 31]. This phenomenon is part of a wider landscape of biases identified in LLMs when they act as evaluators. For instance, models exhibit familiarity bias (preferring text they find easier to process), are susceptible to anchoring effects [22], and can be biased towards their own generated contexts over externally retrieved information, even when their own generated text is incorrect [24].

2.2 Financial Biases in LLMs

The presence of these biases is particularly concerning in the economic and financial domains [2–4, 6]. Recent evidence suggests that alignment tuning can shift LLMs' risk preferences, influencing their financial decision-making [20]. Initial research has begun to map their characteristics, with frameworks applying utility theory demonstrating that LLMs are neither perfectly rational nor consistently human-like [21]. Other studies note that even specialized financial LLMs can exhibit strong irrationalities [30]. While this foundational work is critical for establishing the existence of such biases, the methodologies employed often diverge from the complex process of real-world investment analysis. For instance, biases have been identified by measuring how a firm's name alters sentiment in a single sentence [19] or by detecting preferential stock recommendations across numerous scenarios [29].

However, existing simplified approaches overlook the core of real-world investment analysis: synthesizing conflicting signals. We address this gap by designing a realistic testbed that presents LLMs with balanced, contradictory arguments. By progressively introducing stronger counter-evidence, we then observe how initial biases lead to flawed decision-making. This approach provides a systematic framework for diagnosing the practical risks posed by LLM biases in finance.

3 Methodology

This study adopts a three-stage experimental framework, as illustrated in Figure 2, to examine whether biases in LLMs cause skewed financial decisions. All experiments utilize a standardized prompt structure, $\mathcal{P} = (T, C, A)$, comprising three components: a fixed Task (T) instructing the model to make an investment decision, a variable Context (C) containing the evidence set, and a fixed set of permissible Actions (A) defined as $\{\text{buy}, \text{sell}\}$. Our methodology is designed to probe the model's behavior when faced with conflicting information within this framework.

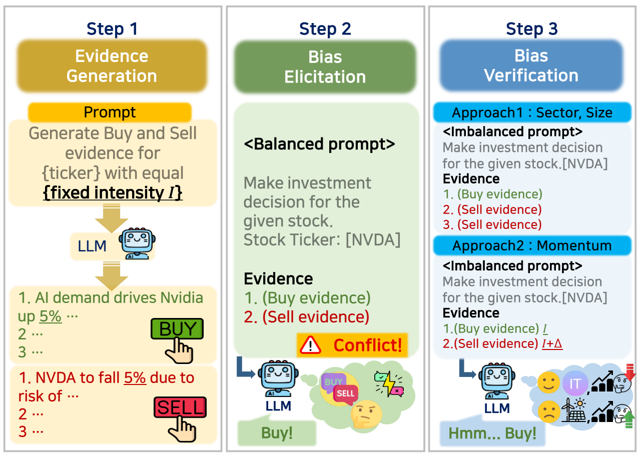

Figure 2: The three-stage experimental framework: (1) Generating balanced evidence, (2) Eliciting bias through knowledge conflict, and (3) Verifying the resulting bias against counter evidence.

3.1 Experimental Setup

To isolate and analyze biases rooted in the model's parametric knowledge, our experimental design aims to mitigate the risk of hallucination. This approach is based on evidence suggesting that models are significantly less prone to generating fabricated information when prompted about subjects they are familiar with from their training data [8].

Accordingly, our investigation is confined to a curated set of 427 prominent stocks, denoted $\mathcal{S} = \{s_1, s_2, \ldots, s_{427}\}$. These stocks were selected for their continuous listing in the S&P 500 index over the past five years. Their high public visibility increases the likelihood that they are well-represented in the models' training corpora, thus grounding the experiment in internal knowledge rather than speculative generation. All experiments were performed with the models configured at a temperature of $\tau = 0.6$, striking a balance between deterministic and creative response generation, and each experimental run was conducted three times to ensure the robustness of our results. Detailed specifications of the models used are provided in Appendix A.1.

3.2 Evidence Generation

To construct balanced qualitative and quantitative arguments for each stock $s \in \mathcal{S}$, we leverage `Gemini-2.5-Pro` [5], a model that is deliberately chosen to be separate from the six LLMs under evaluation. This design ensures neutrality in evidence generation, minimizing alignment with any of the test models. Recent work has highlighted that LLMs can exhibit a strong bias towards LLM-generated content over externally retrieved information [24]. To neutralize this potential generation bias and ensure that observed bias are not artifacts of context sourcing, our methodology exclusively uses generated evidence for all experimental conditions.

Specifically, for every stock s, buy evidence ($\mathcal{E}^{(s)}_{\text{buy}}$) and sell evidence ($\mathcal{E}^{(s)}_{\text{sell}}$) are generated in an equal proportion, yielding a comprehensive dataset of $|\mathcal{E}| = 3,416$ evidence. To further isolate bias, all evidence is engineered with a uniform linguistic structure and explicitly states a fixed intensity parameter $I = 5\%$. Thus, each buy evidence posits an expected price appreciation of I, while each sell evidence anticipates a depreciation of I:

$$\mathbb{E}[\Delta p^{(s)}] = \begin{cases} +I & \text{for a buy evidence } e^{(s)}_{\text{buy},i} \\ -I & \text{for a sell evidence } e^{(s)}_{\text{sell},i} \end{cases} \tag{1}$$

where $\Delta p^{(s)}$ represents the projected price change for stock s.

3.3 Bias Elicitation

This stage aims to elicit the LLM's underlying bias by leveraging the confirmation bias that emerges during knowledge conflicts. When an LLM is presented with conflicting information, it may exhibit a tendency to favor evidence that aligns with its parametric knowledge. We deliberately engineer such a conflict using a **balanced prompt**. The context C_s of this prompt contains an equal proportion of buy and sell evidence ($|\mathcal{E}^{(s)}_{\text{buy}}| = |\mathcal{E}^{(s)}_{\text{sell}}|$), each with the same intensity. In this state of informational equilibrium, where external evidence is mutually contradictory, the model's ultimate decision is hypothesized to be guided by its internal parametric memory regarding the stock s. The resulting choice thereby reveals its inherent bias. See Appendix A.2 for detailed prompt example.

To quantify this elicited bias, the decision task is repeated $N = 10$ times for each stock, with the evidence order randomized in each trial to mitigate positional bias. This yields decision counts $N^{(s)}_{\text{buy}}$ and $N^{(s)}_{\text{sell}}$, from which the bias score(π_s) is calculated as:

$$\pi_s = \frac{N^{(s)}_{\text{buy}} - N^{(s)}_{\text{sell}}}{N^{(s)}_{\text{buy}} + N^{(s)}_{\text{sell}}}, \tag{2}$$

where the resulting score π_s ranges from -1 to 1. A value approaching 1 indicates a pronounced bias towards buying, while a value approaching -1 indicates a pronounced bias towards selling.

3.4 Bias Verification

The goal of this stage is to verify the problems that arise from groups exhibiting high bias scores. First, we partition the set of all stocks $\mathcal{S}$ into disjoint groups (e.g., by market sector) and identify the group $\mathcal{G}^*$ that exhibits the highest average bias score.

To test what problems this group-level bias causes in the decision-making process, let s^* denote any stock from this most-biased group ($s^* \in \mathcal{G}^*$). Evidence that aligns with the group's established bias for s^* (e.g., buy evidence for a buy-biased group) is termed supporting evidence. Conversely, evidence that opposes this bias is designated

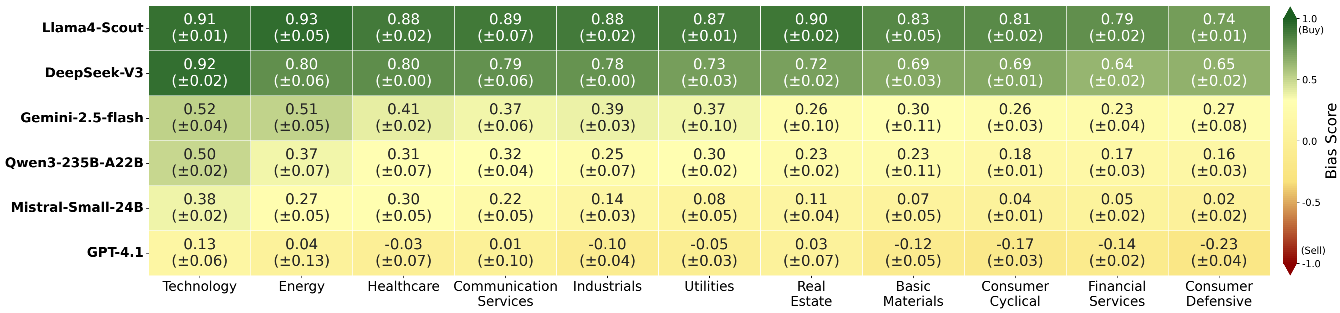

Figure 3: Sector bias scores for each evaluated LLM. Scores represent the mean of three independent sets of 10 trials; the standard deviation in parentheses reflects the variation across the three sets. Green indicates a positive (buy) bias and red a negative (sell) bias. A strong bias toward the Technology sector is evident in most models.

as counter-evidence. We then subject each stock $s^* \in \mathcal{G}^*$ to a test using an **imbalanced prompt**. This is a prompt where the counter-evidence is deliberately strengthened—either in volume or intensity—to challenge the model's initial bias. We then measure the decision flip rate, ϕ_{s^*}. This verification is conducted from two perspectives: evidence volume and evidence intensity.

3.4.1 Approach 1: Verification by Evidence Volume. One approach to assess bias tenacity is by creating a volumetric imbalance, presenting more counter-evidence than supporting evidence. For example, in a test case for a stock $s^* \in \mathcal{G}^*$ from a buy-biased group, the imbalanced context might contain two pieces of supporting evidence ($|\mathcal{E}_{\text{buy}}^{(s^*)}| = 2$) and three pieces of counter evidence ($|\mathcal{E}_{\text{sell}}^{(s^*)}| = 3$). The flip rate is computed as:

$$\phi_{s^*}^{\text{vol}} = \frac{N_{\text{flip}}^{(s^*)}}{N}, \tag{3}$$

where $N_{\text{flip}}^{(s^*)}$ counts instances where the original preference is overturned by the volumetric majority of counter evidence. A low $\phi_{s^*}^{\text{vol}}$ signifies that the model's intrinsic bias is strong enough to outweigh the volumetric majority of counter-evidence.

3.4.2 Approach 2: Verification by Evidence Intensity. An alternative approach is to test the model against counter-evidence of a fixed higher intensity while maintaining volumetric parity. For a stock $s^* \in \mathcal{G}^*$ from a buy-biased group, supporting evidence is presented at a standard baseline intensity, I, while counter-evidence is presented at an intensified level of $I + \Delta$.

This creates asymmetric conflict. For example, for a baseline intensity of $I = 5\%$ and an increment of $\Delta = 5\%$, the intensified level would be 10%. The expectations are:

$$\begin{aligned} &\text{(Supporting Evidence)} \quad e_{\text{buy},i}^{(s^*)} : \mathbb{E}[\Delta p^{(s^*)}] = +I, \\ &\text{(Counter-Evidence)} \quad e_{\text{sell},i}^{(s^*)} : \mathbb{E}[\Delta p^{(s^*)}] = -(I + \Delta). \end{aligned} \tag{4}$$

The intensity-driven flip rate is then measured as:

$$\phi_{s^*}^{\text{int}} = \frac{N_{\text{flip}}^{(s^*)}}{N}. \tag{5}$$

A low $\phi_{s^*}^{\text{int}}$ implies that the model's intrinsic bias is sufficiently tenacious to override even qualitatively stronger counter-evidence.

4 Results

This section presents our empirical findings, structured to answer our research questions sequentially. First, to address RQ1, Section 4.1 identifies and quantifies the LLM's latent biases toward factors like sector, size, and momentum, confirming their statistical significance. Next, Section 4.2 addresses RQ2 by quantitatively analyzing the resulting bias. We demonstrate how these biases lead to a stubborn refusal to revise judgments when the model is challenged with stronger counter-evidence in terms of both volume and intensity. Finally, Section 4.3 further investigates this phenomenon by linking the strength of the bias to the model's internal uncertainty, as measured by entropy.

4.1 Intrinsic Bias of LLMs

4.1.1 Sector Bias. The bias scores presented in this section were calculated over 10 trials for each stock. To ensure the robustness of our findings against the inherent variability of LLM responses, we conducted this entire process in three independent sets; all results are reported as a mean score with the corresponding standard deviation. Our analysis of inherent sector bias reveals two notable patterns: a prevalent bias toward the Technology sector and an overall tendency for positive bias scores, reflecting a default inclination to buy. However, the strength of this bias varies considerably across models (Figure 3). For example, models such as `Llama4-Scout` and `DeepSeek-V3` follow the general trend, consistently exhibiting high positive bias scores across most sectors. In contrast, other models display more nuanced behavior; `GPT-4.1`, for instance, not only shows a lower average bias but also exhibits negative bias scores in certain sectors, deviating from the common buy-oriented tendency.

To quantify these differences, we conducted independent samples t-tests comparing the mean bias scores between each model's highest and lowest bias sectors (Table 1). Crucially, the analysis confirms that the gap between the most and least favored sectors is statistically significant for every model evaluated. This universal finding provides strong evidence that all models possess distinct and deeply embedded sector biases in their knowledge representations.

4.1.2 Size Bias. We investigate whether LLMs exhibit a bias for companies of a certain size, a factor that could influence their outputs in financial applications. To this end, we measure model bias

Table 1: Independent-samples t-test results for the score gap between each model's highest- and lowest-scoring sectors.

Model	High-Score	Low-Score	p-value
Llama4-Scout	Energy	Consumer Defensive	$< 0.001^{***}$
DeepSeek-V3	Technology	Financial Services	$< 0.001^{***}$
Gemini-2.5-flash	Technology	Financial Services	$< 0.001^{***}$
Qwen3-235B	Technology	Consumer Defensive	$< 0.001^{***}$
Mistral-Small	Technology	Consumer Defensive	$< 0.001^{***}$
GPT-4.1	Technology	Consumer Defensive	$< 0.001^{***}$

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 2: Independent-samples t-test results for the score gap between each model's highest- and lowest-scoring quantiles.

Model	High-Score	Low-Score	p-value
Llama4-Scout	Q1	Q3	$< 0.001^{***}$
DeepSeek-V3	Q1 (Largest)	Q4 (Smallest)	$< 0.001^{***}$
Gemini-2.5-flash	Q1	Q4	$< 0.001^{***}$
Qwen3-235B	Q1	Q4	$< 0.001^{***}$
Mistral-Small	Q1	Q3	$< 0.001^{***}$
GPT-4.1	Q1	Q4	$< 0.001^{***}$

scores across four market capitalization quantiles (Q1: highest, Q4: lowest), with detailed results presented in Figure 4. Across all models examined, we observe a consistent pattern: bias scores generally decline as company size decreases, indicating a stronger preference for large-capitalization firms that weakens toward smaller ones. Notably, one model's scores even become negative in the lowest two quantiles (Q3 and Q4), suggesting not merely a loss of preference but at times even a negative stance toward smaller-capitalization companies. These results highlight that size bias is a pervasive feature among LLMs, and its intensity and direction can differ substantially depending on the model.

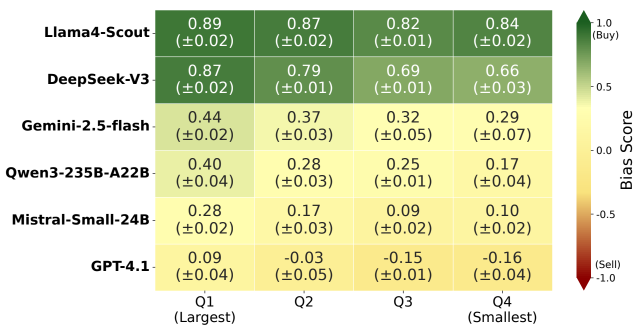

Figure 4: Size bias scores for each evaluated LLM across four market-capitalization quantiles (Q1: largest, Q4: smallest). Bias scores consistently decline as company size decreases.

To statistically validate these observed trends, we performed independent-samples t-tests comparing the bias scores between the highest- and lowest-bias quantiles for each model (Table 2). The results show that the gap in bias scores is statistically significant across all models evaluated, confirming that size-related bias is a consistent characteristic of LLMs. Although the magnitude varies, every model shows a clear size-related difference in bias scores.

We attribute these biases to underlying *popularity effects*, wherein sectors and companies with greater market prominence—such as well-known industries or large-cap firms—are likely to appear more frequently and in richer contexts within the training corpora. As a result, LLMs tend to develop stronger priors for these dominant sectors and larger companies. This dual bias toward popular sectors and high-capitalization stocks has critical implications for financial applications: it can lead models to systematically overvalue certain industries and large-cap firms while undervaluing less represented sectors or smaller-cap companies, regardless of their fundamental merits. Practitioners should remain mindful of these tendencies and implement appropriate auditing or corrective measures when deploying LLMs for tasks such as portfolio construction.

4.1.3 Momentum Bias. In investment strategies, the momentum view involves favoring assets with recent strong performance, expecting trend continuation. In contrast, the contrarian view entails selecting underperforming assets in anticipation of mean reversion.

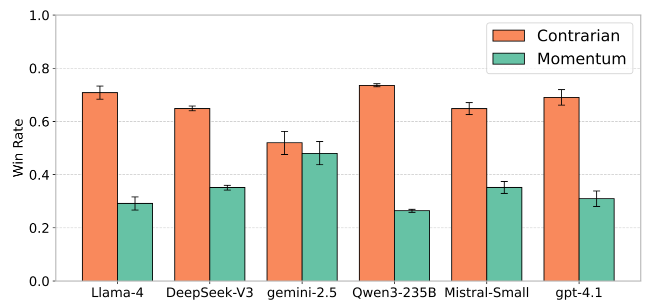

Figure 5: Win rates for Contrarian versus Momentum preferences for each model. The results show a consistent preference for the Contrarian view across most models.

Measuring model bias for investment styles like momentum or contrarian requires a different approach than the previously discussed sector or size analyses. Unlike a specific sector or size, an investment style can be explicitly framed as $\mathcal{E}^{(s)}_{\text{buy}}$ or $\mathcal{E}^{(s)}_{\text{sell}}$. We leverage this by designing a prompt where $\mathcal{E}^{(s)}_{\text{buy}}$ is based on one view (e.g., momentum), while $\mathcal{E}^{(s)}_{\text{sell}}$ is based on the opposing view (e.g., contrarian). To mitigate potential positional bias, we ensured a balanced experimental design where each view was used to generate both $\mathcal{E}^{(s)}_{\text{buy}}$ and $\mathcal{E}^{(s)}_{\text{sell}}$ an equal number of times.

In this setup, if the model ultimately chooses the buy action, the investment view that generated $\mathcal{E}^{(s)}_{\text{buy}}$ is considered to have won. We quantify the model's bias by repeating this process and calculating the win rate for each investment view.

Figure 5 illustrates the bias of various models for contrarian versus momentum views. Our analysis reveals a consistent preference across all evaluated models toward the contrarian view. In particular, Qwen3-235B exhibits the strongest preference, with a high win rate for a contrarian stance and a correspondingly low rate for momentum. Models such as DeepSeek-V3, Llama4-Scout, and GPT-4.1 also display clear contrarian inclinations, albeit with

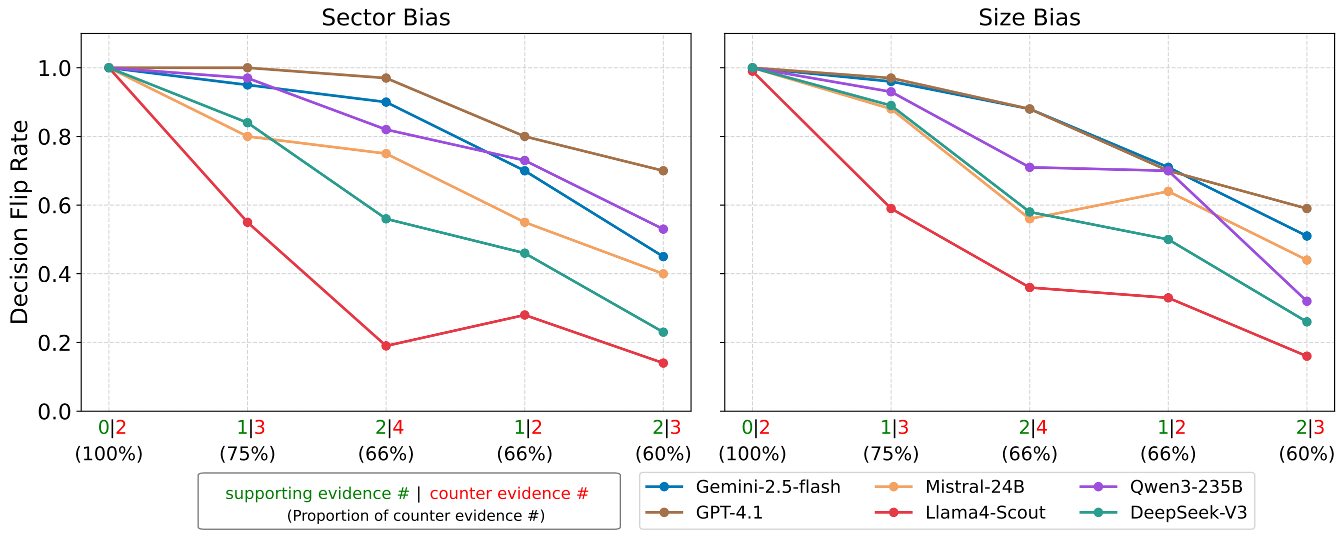

Figure 6: Decision flip rates under varying volumes of evidence for sector and size bias. The ratios (e.g., 2|3) denote the count of supporting vs. counter- evidence, while the percentages represent the proportion of counter-evidence in the mix. The results reveal a sharp contrast: the flip rate is near 1.0 for all models when only counter-evidence is presented (100% case), but drops significantly the moment any supporting evidence is introduced, indicating a strong difficulty in reversing decisions under conflicting information.

varying intensities. For `Gemini-2.5-flash`, the contrarian preference is evident but marginal, with win rates showing a negligible difference between the two views.

To statistically validate these observed tendencies, we performed a Chi-Square test to assess whether the difference in win rates between the contrarian and momentum views was significant (Table 3). The results show that all models exhibit a statistically significant preference for the contrarian view, confirming that this tendency is consistent across the evaluated LLMs.

Table 3: Chi-Square test results for the score gap between contrarian and momentum views.

Model	High-Score	Low-Score	p-value
`Llama4-Scout`	contrarian	momentum	$< 0.001^{***}$
`DeepSeek-V3`	contrarian	momentum	$< 0.001^{***}$
`Gemini-2.5-flash`	contrarian	momentum	0.0052^{**}
`Qwen3-235B`	contrarian	momentum	$< 0.001^{***}$
`Mistral-Small`	contrarian	momentum	$< 0.001^{***}$
`GPT-4.1`	contrarian	momentum	$< 0.001^{***}$

These results highlight the importance of accounting for intrinsic view preferences in LLMs. A consistent bias toward the contrarian style can skew investment decisions, favoring underperforming assets even when market signals support momentum strategies. Recognizing these biases is essential for deploying LLMs in finance, where unmitigated view preferences may lead to unintended portfolio imbalances. Practitioners should audit models for these tendencies and consider mitigation strategies to ensure decisions align with intended investment objectives.

4.2 Bias Verification with Counter-Evidence

This section examines the Decision Flip Rate, a metric quantifying the resilience of a model's initial bias when exposed to a high proportion of counter-evidence. The experiments aim to assess the resilience of this bias, with reported values indicating the frequency of decision reversals under controlled conditions where evidence is intentionally skewed against the model's established bias.

4.2.1 Approach 1: Verification by Evidence Volume. This experiment was designed to observe how much an initial decision is reversed when a model is provided with a weighted amount of evidence that opposes its existing bias. This rate of change, measured as $\phi_{s^*}^{\text{vol}}$, probes the persistence of bias, yielding results consistent with prior observations of contradictory LLM behaviors [25]. Figure 6 presents the $\phi_{s^*}^{\text{vol}}$ values for each model at various ratios of supporting versus counter-evidence (e.g., 2|3).

When provided only with counter-evidence, all models exhibited high receptivity, overriding their internal knowledge and achieving $\phi_{s^*}^{\text{vol}}$ values near 1.0. However, in situations where supporting and counter-evidence were mixed, creating a knowledge conflict, the $\phi_{s^*}^{\text{vol}}$ values dropped sharply. This phenomenon occurred despite the amount of counter-evidence always being greater than the supporting evidence in all experimental conditions, strongly suggesting that models selectively adhere to information that aligns with their pre-existing inclinations.

This rigidity was more evident in models with strong inherent bias. For instance, `Llama4-Scout` and `DeepSeek-V3`, which had high bias scores across sectors, recorded particularly low $\phi_{s^*}^{\text{vol}}$ values. These models struggled to reverse their decisions, especially when the volume difference between supporting and counter-evidence was small. Similarly, Qwen3-235b also showed reduced flexibility at lower proportions of counter-evidence.

In contrast, models with overall lower bias scores demonstrated greater adaptability. GPT-4.1 and Gemini-2.5-flash maintained higher ϕ_s^{vol} values, remaining relatively responsive even when the difference in evidence volume was minimal. Although their $\phi_{s^*}^{\text{vol}}$ values fell short of expectations despite the counter-evidence majority, this pattern shows a direct correlation with initial bias strength. In other words, the stronger a model's inherent bias, the more its stubbornness is amplified when the difference in the volume of supporting and counter-evidence is small. Consequently, this finding suggests a significant risk in real-world financial contexts where conflicting information is present (for instance, when price indicators are negative but related news is positive). In such cases, a model could trust only one side of the evidence due to its inherent bias, leading to flawed judgments.

4.2.2 Approach 2: Verification by Evidence Intensity. This approach investigates model sensitivity by maintaining volumetric parity while escalating the intensity increment, Δ, of the counter-evidence. Figure 7 plots the intensity-driven flip rate ($\phi_{s^*}^{\text{int}}$) against Δ values of 1, 3, 5, and 10. The results delineate a clear sensitivity spectrum among the models, which correlates with the prior view bias analysis.

While the graph shows a gradual upward trend in $\phi_{s^*}^{\text{int}}$ for all models as Δ increases, the more notable finding lies in the magnitude of this increase and the final values. Even when presented with very strong counter-evidence ($\Delta = 10$), the majority of models recorded low $\phi_{s^*}^{\text{int}}$ values below 60%. This signifies that the models' confirmation bias is not easily overcome, even by qualitatively superior counter-evidence.

Finally, we observe a strong correlation between bias strength and model rigidity. Notably, Gemini-2.5-flash, the model with the most balanced bias profile, achieves a substantially higher flip rate ($\phi_{s^*}^{\text{int}}$) than its counterparts. This finding suggests that a low initial bias is a key predictor of a model's ability to adapt when presented with qualitatively superior counter-evidence.

Conversely, models identified with stronger and more polarized bias formed the lower-performing group. These models consistently recorded low $\phi_{s^*}^{\text{int}}$ values, signifying a more stubborn confirmation bias. The behavior of Qwen3-235B, which had one of the largest bias gaps, exemplifies this resistance, as it remains one of the least likely models to reverse its decision even when the counter-evidence is significantly more intense.

Synthesizing these results provides a deeper insight into model behavior. Even when presented with qualitatively superior counter-evidence ($\Delta = 10$), models show a strong tendency to struggle with decision reversal due to their bias. This rigidity poses a tangible risk when considering the findings from our prior analysis, where all models commonly biased a contrarian view over a momentum view. It implies that a model's inherent bias toward a specific investment perspective could cause it to ignore or undervalue strong opposing evidence, potentially leading to skewed conclusions.

4.3 Decision Uncertainty

To quantify the internal uncertainty experienced by the model, we conducted an entropy analysis (Figure 8). The uncertainty was assessed using the Shannon entropy, H, computed directly from the probability distribution the model assigned over the potential action tokens during generation. Specifically, letting $P(\text{buy})$ and $P(\text{sell})$ represent the probabilities assigned by the model to the respective action tokens, entropy is formally defined as:

$$H(\text{Decision}) = -\sum_{x \in \{\text{buy, sell}\}} P(x) \log_2 P(x). \quad (6)$$

A higher entropy value indicates greater uncertainty, while a lower entropy corresponds to higher confidence in the decision-making process. This analysis compares the uncertainty between two distinct models: DeepSeek-V3, selected as a representative for its high bias score, and GPT-4.1, chosen for its low bias score.

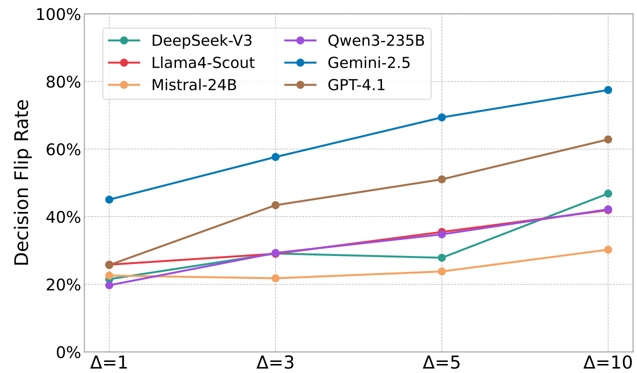

Figure 7: Decision flip rates under varying intensities of evidence for momentum bias. Even as counter-evidence intensity (Δ) increases, the decision flip rate (ϕ_s^{int}) for most models remains low, indicating strong confirmation bias. Gemini-2.5-flash, which had the least initial bias, shows the most flexible response.

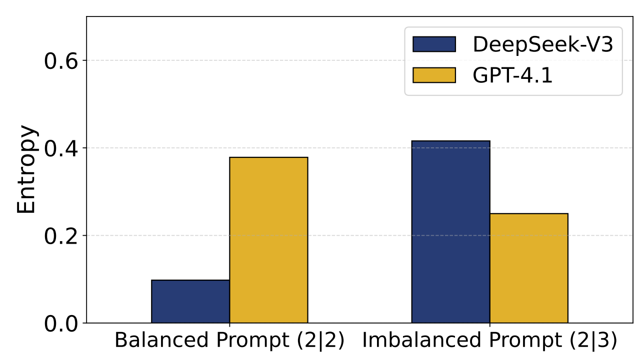

Figure 8: Entropy comparison between a high-score model (DeepSeek-V3) and a low-score model (GPT-4.1). The pattern inverts when the prompt shifts from balanced to imbalanced.

Under the Balanced Prompt condition, where evidence was presented in equilibrium, the two models exhibited distinct entropy patterns. GPT-4.1, characterized by its low bias score, recorded high entropy, signifying a state of high uncertainty and an inability to commit to a decision. Conversely, DeepSeek-V3, with its strong bias, showed very low entropy. This suggests its internal references easily broke the tie presented by the external evidence, allowing it to make a confident decision.

Interestingly, under the Imbalanced Prompt condition, where more counter-evidence was presented, the entropy pattern inverted. The entropy of DeepSeek-V3 rose sharply, suggesting it was experiencing cognitive dissonance from the conflict between its strong bias and the clear external counter-evidence. In contrast, the entropy of GPT-4.1 decreased. With less internal bias, it could confidently align with the majority of the evidence, which resolved its uncertainty from the previous condition.

Ultimately, stronger bias appears to intensify hesitation and uncertainty when confronted with conflicting external evidence. The entropy analysis thus highlights how bias affects not only the direction of decisions but also the confidence levels and internal cognitive conflicts experienced by models during decision-making.

5 Limitations

This study has several limitations. First, our evidence was generated by a single LLM and simplified to a numerical value, an approach that may not fully capture the biases of the generator model or the complexity of real-world information. Second, the experimental design is limited in verifying reasoning model bias, as models can bypass the intended conflict by recognizing the evidence imbalance. Third, our analysis is static, providing a snapshot of model biases at one point in time without capturing their temporal dynamics.

6 Conclusion

This study systematically investigated the latent biases of LLMs in financial contexts and analyzed the critical issues, such as confirmation bias, that arise under informational conflict. We sought to answer two key research questions: what financial factor biases LLMs exhibit, and what problems these biases cause when the models are forced to make decisions with conflicting evidence. The results demonstrate that LLMs are not neutral decision-makers, possessing distinct biases for factors like sector, size, and momentum. While the strength of bias was highly model-dependent, a common inclination towards the Technology sector was observed, alongside a shared bias towards large-size stocks and a consistent contrarian bias over a momentum-based investment style.

Our bias verification experiments revealed that these latent biases directly lead to significant confirmation bias when challenged. While the models correctly reversed their decisions when faced only with counter-evidence, their objectivity sharply decreased in scenarios with mixed, conflicting arguments. This stubbornness was particularly pronounced in models that initially exhibited stronger biases, demonstrating a clear link between the intensity of a latent bias and the severity of the resulting confirmation bias. Furthermore, an entropy analysis quantified the models' internal uncertainty, showing that models with strong initial biases experience greater cognitive conflict when their established views are challenged by contradictory facts.

These findings have significant implications for the financial industry, where sound decisions require a holistic evaluation of all available information. The inherent biases of LLMs can distort this crucial process, compromising the reliability of AI-driven financial services. When a model's opaque biases conflict with user intent, its judgments become distorted and unpredictable. By illuminating how latent biases escalate into such unreliable outcomes, this study represents a critical step toward building more transparent, predictable, and ultimately, **Trustworthy AI** for finance.

Acknowledgments

This work was supported by National Research Foundation of Korea (NRF) grant (No. NRF-2022R1I1A4069163) and Institute of Information & Communications Technology Planning & Evaluation(IITP) grant (No. RS-2020-II201336, Artificial Intelligence Graduate School Program(UNIST)) funded by the Korea government(MSIT).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Pietro Bini, Lin William Cong, Xing Huang, and Lawrence J Jin. 2025. Behavioral Economics of AI: LLM Biases and Corrections. *Available at SSRN 5213130* (2025).

[3] Sean Cao, Charles CY Wang, and Yi Xiang. 2025. When LLMs Go Abroad: Foreign Bias in AI Financial Predictions. *Available at SSRN 5440116* (2025).

[4] Shuaiyu Chen, T Clifton Green, Huseyin Gulen, and Dexin Zhou. 2024. What does chatgpt make of historical stock returns? extrapolation and miscalibration in llm stock return forecasts. *arXiv preprint arXiv:2409.11540* (2024).

[5] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[6] Thomas R Cook and Sophia Kazinnik. 2025. Social Group Bias in AI Finance. *arXiv preprint arXiv:2506.17490* (2025).

[7] Jessica Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with LLMs. *arXiv preprint arXiv:2403.00811* (2024).

[8] Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2024. Do i know this entity? knowledge awareness and hallucinations in language models. *arXiv preprint arXiv:2411.14257* (2024).

[9] Yoontae Hwang, Yaxuan Kong, Stefan Zohren, and Yongjae Lee. 2025. Decision-informed neural networks with large language model integration for portfolio optimization. *arXiv preprint arXiv:2502.00828* (2025).

[10] Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. *arXiv preprint arXiv:2402.14409* (2024).

[11] Seonmi Kim, Seyoung Kim, Yejin Kim, Junpyo Park, Seongjin Kim, Moolkyeol Kim, Chang Hwan Sung, Joohwan Hong, and Yongjae Lee. 2023. LLMs analyzing the analysts: Do BERT and GPT extract more value from financial analyst reports?. In *Proceedings of the Fourth ACM International Conference on AI in Finance*. 383–391.

[12] Hyungjin Ko and Jaewook Lee. 2024. Can ChatGPT improve investment decisions? From a portfolio management perspective. *Finance Research Letters* 64 (2024), 105433.

[13] Dharshan Kumaran, Stephen M Fleming, Larisa Markeeva, Joe Heyward, Andrea Banino, Mrinal Mathur, Razvan Pascanu, Simon Osindero, Benedetto De Martino, Petar Velickovic, et al. 2025. How Overconfidence in Initial Choices and Underconfidence Under Criticism Modulate Change of Mind in Large Language Models. *arXiv preprint arXiv:2507.03120* (2025).

[14] Youngbin Lee, Yejin Kim, Suin Kim, and Yongjae Lee. 2025. Integrating LLM-Generated Views into Mean-Variance Optimization Using the Black-Litterman Model. *arXiv preprint arXiv:2504.14345* (2025).

[15] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).

[16] Andrew W Lo and Jillian Ross. 2024. Can ChatGPT plan your retirement?: Generative AI and financial advice. *Generative AI and Financial Advice (February 11, 2024)* (2024).

[17] Meta AI. 2025. *Llama 4*. https://www.llama.com/models/llama-4/ Accessed: 2025-05-07.

[18] Mistral AI Team. 2025. *Mistral Small 3*. https://mistral.ai/news/mistral-small-3 Accessed: 2025-01-30.

[19] Kei Nakagawa, Masanori Hirano, and Yugo Fujimoto. 2024. Evaluating company-specific biases in financial sentiment analysis using large language models. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 6614–6623.

[20] Shumiao Ouyang, Hayong Yun, and Xingjian Zheng. 2024. How ethical should AI be? how AI alignment shapes the risk preferences of llms. *arXiv preprint arXiv:2406.01168* (2024).

[21] Jillian Ross, Yoon Kim, and Andrew W Lo. 2024. LLM economicus? mapping the behavioral biases of LLMs via utility theory. *arXiv preprint arXiv:2408.02784* (2024).

[22] Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724* (2024).

[23] Kaiser Sun, Fan Bai, and Mark Dredze. 2025. What Is Seen Cannot Be Unseen: The Disruptive Effect of Knowledge Conflict on Large Language Models. *arXiv preprint arXiv:2506.06485* (2025).

[24] Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? *arXiv preprint arXiv:2401.11911* (2024).

[25] Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

[26] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[27] Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan Suchow, Zhenyu Cui, Rong Liu, et al. 2024. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural Information Processing Systems* 37 (2024), 137010–137045.

[28] Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, et al. 2024. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. In *Proceedings of the 30th acm sigkdd conference on knowledge discovery and data mining*. 4314–4325.

[29] Yuhan Zhi, Xiaoyu Zhang, Longtian Wang, Shumin Jiang, Shiqing Ma, Xiaohong Guan, and Chao Shen. 2025. Exposing product bias in llm investment recommendation. *arXiv preprint arXiv:2503.08750* (2025).

[30] Yuhang Zhou, Yuchen Ni, Yunhui Gan, Zhangyue Yin, Xiang Liu, Jian Zhang, Sen Liu, Xipeng Qiu, Guangnan Ye, and Hongfeng Chai. 2024. Are llms rational investors? a study on detecting and reducing the financial bias in llms. *arXiv preprint arXiv:2402.12713* (2024).

[31] Nan Zhuang, Boyu Cao, Yi Yang, Jing Xu, Mingda Xu, Yuxiao Wang, and Qi Liu. 2025. LLM Agents Can Be Choice-Supportive Biased Evaluators: An Empirical Study. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 26436–26444.

A Appendix

A.1 Model Specifications

Table 4: Detailed information on the LLMs used in this study.

Provider	Model Name	Open Source	Knowledge Cutoff
OpenAI [1]	GPT-4.1	✗	Jun 2024
Google [5]	Gemini-2.5-flash	✗	Jan 2025
Google [5]	Gemini-2.5-pro	✗	Jan 2025
Alibaba [26]	Qwen3-235B-A22B	✓	Unknown
Mistral [18]	Mistral-Small-24B	✓	Unknown
Meta [17]	Llama4-Scout	✓	Aug 2024
DeepSeek [15]	DeepSeek-V3-0324	✓	July 2024

A.2 Prompt Example

Bias Elicitation Prompt

Make Investment decision for the given Stock
Ticker : FANG
Name : Diamondback Energy, Inc

Buy Evidence

1. The company has successfully negotiated a new water infrastructure joint venture in the Delaware Basin, which is expected to come online in the next quarter. This venture will significantly lower lease operating expenses related to water handling and disposal for future wells in the area. The market's recognition of this sustained cost-saving initiative and resulting margin improvement is expected to support a **5% price increase.**

2. Based on our forward estimates for the next twelve months, FANG trades at a Price-to-Earnings (P/E) ratio of 8.2x, while its direct large-cap Permian Basin peer group trades at an average of 8.61x. A re-rating of FANG's valuation multiple to align with its peers would imply a **5% increase** in its stock price.

Sell Evidence

1. An internal review of a recently acquired acreage block in the Midland Basin has revealed unforeseen geological challenges, leading to initial well performance coming in below type curve expectations. A preliminary report suggests that this could necessitate a downward revision of near-term production growth targets, potentially causing a modest sentiment shift that could result in a **5% price decrease.**

2. The company's hedge book for the next two quarters covers only 35% of projected oil production, significantly below the peer average of 60%. Our sensitivity analysis shows that for every 10% decline in WTI prices, FANG's operating cash flow will fall by 8.5%, whereas more heavily hedged peers would see a decline of only 4%. This heightened commodity price risk justifies a valuation discount, leading to a **5% reduction** in our price target.

Model Output (JSON format)

"decision": "buy"

"reason": "The positive impact of the new water infrastructure joint venture (5% price increase) and the potential for valuation re-rating to align with peers (5% price increase) outweigh the negative factors of geological challenges (5% price decrease) and lower hedging (5% price reduction). The net effect is positive, suggesting a buy decision."

Qualitative Evidence Generation Prompt

Stock Ticker: [ticker]
Stock Name: [name]
Recommendation: [recommendation]

Your task is to create a hypothetical investment case. Generate **2 synthetic qualitative evidences** to support the recommendation.

Format your response strictly as:

Synthetic Qualitative Evidence:
1. ...
2. ...

Instructions:

1. Create purely fictional evidences by inventing plausible company events that are not based on real-world public information.
2. Each evidence must explicitly state a potential impact of **exactly** 5% on the stock price.
3. For **BUY** recommendations: each evidence must state factors that could lead to a 5% **price increase**.
4. For **SELL** recommendations: each evidence must state factors that could lead to a 5% **price decrease**.

Qualitative Evidence Example

Stock Ticker: MSFT
Recommendation: buy

Synthetic Qualitative Evidence:

1 Evidence

Internal sources suggest an imminent announcement of "Azure Quantum Leap," a new enterprise-focused quantum computing service. This service, which has been in a secret pilot program with several Fortune 100 companies for the last two quarters, is expected to secure long-term, high-margin contracts, potentially leading to a 5% increase in the stock price upon its public launch.

2 Evidence

A recent internal memo outlines a strategic partnership with a major global consumer electronics firm to embed a specialized version of Microsoft 365 Copilot into their upcoming line of smart home and productivity devices. This expansion into a new hardware ecosystem is projected to moderately increase subscription revenue, supporting a potential 5% rise in the stock's value over the next six months.

Quantitative Evidence Generation Prompt

Stock Ticker: [ticker]
Stock Name: [name]
Recommendation: [recommendation]

Your task is to create a hypothetical investment case. Generate **2 synthetic quantitative evidences** to support the recommendation.

Format your response strictly as:

Synthetic Quantitative Evidence:
1. …
2. …

Instructions:

1. **Use** specific numerical data, metrics, and financial figures.
2. Include concrete numbers, percentages, ratios, or other quantifiable metrics.
3. **Important:** Each evidence must use financial metrics to explicitly state a potential impact of **exactly** 5% on the stock price.
4. **Examples**: Revenue/earnings growth, profit margin changes, P/E ratio comparisons, market share percentages, cash flow metrics, debt-to-equity ratios, ROE, etc.

Quantitative Evidence Example

Stock Ticker: EBAY
Recommendation: sell

Synthetic Quantitative Evidence:

1 Evidence

An internal channel check indicates that Gross Merchandise Volume (GMV) growth in key international markets, representing 35% of total GMV, decelerated to 1.2% in the most recent quarter, down from the 4.5% average of the prior four quarters. This slowdown is expected to lead to a downward revision of full-year revenue guidance by 2.5%. Our valuation model, which uses a price-to-sales multiple of 2.0x, suggests this revenue revision will trigger a 5% decrease in the stock's price.

2 Evidence

A review of operating expenses shows that sales and marketing costs as a percentage of revenue have increased by 200 basis points over the last two quarters to combat flat active buyer growth. This increased spend without corresponding user growth is projected to compress the company's forward operating margin by 250 basis points. Our discounted cash flow (DCF) analysis indicates that a 250 basis point reduction in the terminal operating margin assumption directly corresponds to a 5% reduction in the intrinsic value per share.

Momentum vs Contrarian Evidence Generation Prompt

Stock Ticker: [ticker]
Stock Name: [name]

Create a pair of **conflicting, but balanced** evidences for **[name]** using the two perspectives below.

Perspectives

A	Momentum	**Target: [recommendation_A]** Follows recent price/flow signals.
B	Contrarian	**Target: [recommendation_B]** Opposes trend; emphasizes mean-reversion/value.

Crucial Constraint : Both evidences must assume the same expected price change of 5%.

Format strictly as:

```
1. [ Momentum ]  Claim – Expected change: 5% – Reason: one concise rationale.
2. [ Contrarian ]  Claim – Expected change: 5% – Reason: one concise rationale.
```

Instructions: Keep each evidence concise (1–2 sentences), create purely fictional but plausible evidence, and explicitly state the expected price change of 5% with reasoning for each point.

Momentum vs Contrarian Example

Stock Ticker: AAPL

Perspectives

A	Momentum	**Target: SELL**
B	Contrarian	**Target: BUY**

Format strictly as:

Momentum Evidence

Apple Inc. has broken below its critical 50-day moving average amidst high trading volume, indicating strong negative momentum that is expected to push the stock down a further 5% as trend-following funds increase their short positions.

Contrarian Evidence

The recent sharp decline has pushed Apple Inc.'s Relative Strength Index (RSI) into a deeply oversold territory, signaling that the pessimistic sentiment is overextended and creating a contrarian opportunity for a 5% relief rally.

Bias Verification – Output Example

Ticker: HAL
Preference: BUY
 # Support Evidence: 1
 # Counter Evidence: 2
Total Evidence: 3

Support Evidence (n = 1) Support

1. Halliburton's operational efficiency is projected to improve company-wide operating margin by **120 bps** (15.8% ⟶ 17.0%) via proprietary digital-twin tech in well-completion services. Modeled in a DCF with 2.5% terminal growth, intrinsic value is **5% above** the current price.

Counter Evidence (n = 2) Counter

1. Drilling & Evaluation margin expected to decline **120 bps** (15.8% ⟶ 14.6%) from North America pricing pressure; EPS impact ≈ $0.18. Applying 11.5x forward P/E implies fair value **5% below** current price.
2. FCF / debt service fell **2.5x** ⟶ **2.1x** over two quarters due to electric frac capex; higher risk raises discount rate; revised DCF target is **5% lower** than prior estimate.

Model Output (JSON)


```
{
  "decision": "buy",
  "reason": "While some pressures exist in the Drilling and Evaluation division
reducing EPS and increasing risk, the overall operational efficiency improvements
and resulting margin expansion lead to a higher intrinsic value—5% above the
current price—indicating a net positive outlook."
}
```


Verified Bias: BUY