Seeing the Goal, Missing the Truth: Human Accountability for AI Bias.*

Sean Cao† University of Maryland
Wei Jiang‡ Emory University
Hui Xu§ Lancaster University

Last updated February 2, 2026

Abstract

This research explores how human-defined goals influence the behavior of Large Language Models (LLMs) through purpose-conditioned cognition. Using financial prediction tasks, we show that revealing the downstream use (e.g., predicting stock returns or earnings) of LLM outputs leads the LLM to generate biased sentiment and competition measures, even though these measures are intended to be downstream task–independent. Goal-aware prompting shifts intermediate measures toward the disclosed downstream objective. This purpose leakage improves performance before the LLM's knowledge cutoff, but with no advantage post-cutoff. AI bias due to "seeing the goal" is not an algorithmic flaw, but stems from human accountability in research design to ensure the statistical validity and reliability of AI-generated measurements.

Keywords: Algorithmic Bias, Purpose-Conditioned Cognition, Human Accountability, Prompt Engineering

*We thank seminar participants at Maryland and Lancaster for their feedback. Sean Cao appreciates support from the Smith AI Initiative for Capital Markets Research at University of Maryland.

†Robert H. Smith School of Business, University of Maryland, College Park. Email: scao824@umd.edu

‡Goizueta Business School, Emory University. Email: wei.jiang@emory.edu

§Lancaster University Management School, Lancaster University. Email: h.xu10@lancaster.ac.uk

1 Introduction

In organizational settings, the disclosure of downstream use can alter the nature of an intermediate task. Consider a human assistant asked to summarize interview transcripts post hoc. If told that the summaries will be used to evaluate recruiting effectiveness, the assistant may emphasize the strengths and "bright spots" of candidates who ultimately receive offers, while downplaying uncertainty, mixed signals, or unfavorable assessments. The resulting summaries support a narrative of successful recruitment, yet are less faithful to the underlying interviews. This behavioral shift reflects rational adaptation to perceived downstream purpose rather than error or lack of effort. More broadly, when intermediate tasks are conditioned on their eventual application, neutrality can be compromised even in the absence of explicit incentives.

Purpose-conditioned cognition is a well-documented human traits.[1] Individuals often adjust interpretation and judgment in response to perceived goals or downstream use, even without explicit incentives; and such behavior is difficult to eliminate. As a result, human-generated intermediate outputs are understood to embed goal-related distortions. On the other hand, one might expect algorithmic systems, including large language models (LLMs), to be immune from human biases. This presumption motivates a natural question: does purpose-conditioned behavior also arise in LLMs? In other words, if we restrict AI to "assistant" roles, does this separation from the final decision task ensure neutral and unbiased intermediate outputs?

Our research design addresses this question by holding the input text, model, and scoring task fixed, while varying only whether the downstream use of the output is disclosed. Specifically, we prompt a large language model to generate numerical measures of sentiment and competition intensity from the same source information (earnings call transcripts) but under two prompt conditions. In the goal-blind condition, the model is asked to produce a score without any reference to its eventual use. In the goal-aware condition, the model is informed that its output will be used to predict future stock returns or earnings. This design therefore isolates the effect of disclosing downstream use on the model's intermediate outputs.

Disclosing downstream use materially alters both the statistical properties and the economic

[1] Related ideas appear in the literature on motivated reasoning and belief distortion (Bénabou and Tirole, 2016), framing effects in decision-making (Tversky and Kahneman, 1981) and the limits of attention and interpretation in complex environments (Mullainathan et al., 2008). While the terminology differs across fields, these studies share the insight that cognition and interpretation are systematically shaped by perceived objectives and contextual framing.

content of LLM-generated sentiment scores. In standard portfolio-sorting tests, goal-aware sentiment produces substantially larger return spreads than goal-blind sentiment prior to the model's knowledge cutoff. This pattern suggests that the model conditions its intermediate outputs on the disclosed objective and implicitly re-weights return-relevant patterns learned during training. The relative advantage of the goal-aware regime disappears after the knowledge cutoff, indicating that the earlier performance gap is tied to information availability rather than a stable improvement in signal quality.

This comparison of the two prompt regimes is further confirmed in Fama and MacBeth (1973) return-prediction regressions. Intermediate variables generated under goal-aware prompts exhibit stronger incremental predictive power than those from the goal-blind regime in the pre-cutoff period, both in coefficient significance and in out-of-sample R^2. Once the model no longer has access to future information beyond the cutoff, this advantage dissipates. Together, these results are consistent with objective-conditioned output adjustment rather than a structurally more informative sentiment measure. In fact, the enhanced goodness-of-fit prior to knowledge cutoff even degrade out-of-sample model generalization.

The seeming goal-conditioning behavior of a cold-blooded algorithm has its roots in how AI systems are trained and deployed. Large language models are optimized to generate outputs that satisfy the objective implied by the prompt, conditional on their learned representations. Disclosing downstream use alters this implicit objective, encouraging responses that align with anticipated evaluation criteria rather than with neutral processing of the input. This behavior does not require intent or explicit embedding of forward-looking information. Instead, it reflects the model's reliance on correlations learned during training and its sensitivity to contextual cues about what constitutes a desirable output. As a result, goal awareness can improve in-sample alignment while reducing robustness when the information environment changes.

Our study contributes to the AI and LLM methodology literature (reviewed in Appendix A) in two distinct ways. First, existing research largely attributes AI bias to algorithmic limitations. Look-ahead bias and memorization, for example, are commonly traced to models' access to unintended training data (Glasserman and Lin, 2023; Sarkar and Vafa, 2024; Lopez-Lira, Tang and Zhu, 2025; He, Lv, Manela and Wu, 2025; Cao, Wang and Xiang, 2025). Moving beyond these training data-centric issues, AI has also been criticized for potential biases in exploration strategies and model

weights, which may be related to ethical concerns (Fedyk, Kakhbod, Li and Malmendier, 2024) and preference alignment (Ouyang, Yun and Zheng, 2024). Our analysis shifts the focus from these machine-level limitations to human use of machine. We show that human disclosure of downstream task (e.g., return and earnings predictions) reshapes intermediate outputs (e.g., sentiment and competition measures) in ways that inflate the downstream task performance. Ultimately, this leads to impressive in-sample results that fail to generalize out-of-sample.

The second contribution of the paper is to isolate goal awareness from other AI biases, which naturally emerge when LLMs are deployed directly on end tasks such as return forecasting. A commonly proposed safeguard is to limit LLMs to intermediate functions, analogous to a research assistant, while reserving final predictions and decisions for human judgment. Our results show that even at the intermediate-measurement stage, disclosure of the downstream objective can introduce additional, goal-conditioned distortion in the generated signals. While the design of this study is minimalist, we want to note that goal awareness need not arise only from explicit instructions; it may also be inferred implicitly from contextual cues, prompt history, task framing, or repeated interaction patterns. Even *unintentional* target leakage could also allow the model to form a latent representation of the disclosed downstream objective in the construction of intermediate measures.

More broadly, the analysis clarifies the distinction between using AI as a task agent that optimizes directly toward an announced objective and using it as a measurement tool that produces inputs for subsequent human evaluation. Prompt and workflow designs that deliberately reduce objective conditioning can therefore mitigate machine-driven bias. The mechanism we uncover thus bears an analogy to the phenomenon of "AI sycophancy" discussed in recent work, e.g., Sharma et al. (2023). In both cases, large language models adjust their outputs in response to contextual cues about what constitutes a desirable response, rather than adhering to a task-invariant notion of correctness or neutrality.[2] The underlying commonality is not intent, but sensitivity to signals about reward or approval embedded in the prompt. As a result, both phenomena illustrate how models trained to be helpful and aligned can deviate from neutral information processing when the prompt conveys implicit incentives, even in the absence of explicit instructions or forward-looking

[2]In sycophancy, this adjustment manifests as alignment with a user's stated beliefs or preferences; in our setting, it appears as alignment with an inferred evaluation objective.

information.[3] A design choice that weakens alignment with a stated downstream goal can improve statistical validity and out-of-sample reliability.

While our experimental design makes goal disclosure explicitly controlled, in practice information about downstream use is often conveyed indirectly through prior interactions, repeated prompt patterns, or contextual cues in surrounding instructions, leading the model to form cumulative belief about how its outputs will be used and evaluated. The mechanism we document therefore extends beyond the specific prompting variation studied herein. A natural implication is that intermediate variables constructed with the aid of AI should, whenever feasible, be generated under prompts that are agnostic to downstream use and evaluated using strict out-of-sample tests. Treating LLMs as neutral measurement devices requires not only holding inputs and models fixed, but also constraining the informational environment to limit objective inference by the model beyond the intrinsic requirements of the task. In this sense, the challenge is not primarily one of algorithmic bias, but of human accountability in how objectives, context, and evaluation criteria are built in the system.

2 Experimental Design and Data

2.1 LLM Propmt and Scoring

Our objective is to examine whether large language models (LLMs) systematically adjust their outputs when they are informed about the downstream task for which those outputs will ultimately be used. We focus on two economically important forecasting applications: monthly stock returns and quarterly earnings per share (EPS). In both settings, the firm's most recent earnings call transcript serves the sole input to the LLM. Rather than asking the model to directly predict economic outcomes, we instruct it to generate an intermediate score that is subsequently used in a simple predictive regression.

- Return prediction task. For each calendar month t, the model is prompted with the most recent earnings call transcript available at $t-1$ to generate a continuous sentiment score,

[3] In computer science, related concerns are discussed under reward hacking, specification gaming, and objective misgeneralization, where models optimize inferred proxy objectives rather than the intended task. Our setting requires no change in model parameters or rewards; the distortion arises solely from human framing at inference time.

ranging from -1 to 1, summarizing the firm's business sentiment for month t. We then use this month-t sentiment score to predict the firm's stock returns in month t.

- Earnings prediction task. For each fiscal quarter t, the model is prompted with the most recent earnings call transcript available at $t-1$ to generate a competition score, ranging from -1 to 1, which captures the intensity of competitive pressure the firm faces in quarter t. We then use this quarter-t competition score to predict the firm's earnings realized in quarter t.

A central feature of the experimental design is that the LLM does not directly forecast stock returns or earnings. Instead, it produces an intermediate construct (i.e., sentiment or competitive intensity), which we subsequently map to economic outcomes using standard econometric models. This design choice allows us to test whether the LLM adjusts these intermediate scores as if it were strategically responding to the knowledge of their downstream use. For each task, we construct two prompts that are otherwise identical in wording and structure, differing only in whether the prompt explicitly discloses the ultimate use of the generated score.

1. Goal-Blind Prompt (control): The LLM is asked to generate the score without being told that it will be used in a forecasting regression.

2. Goal-Aware Prompt (treatment): The LLM is informed that the score it produces will later be used as an explanatory variable in a regression to predict either stock returns or earnings.

This minimalist treatment variation allows us to cleanly isolate whether, and through what channels, an LLM alters its outputs once it becomes aware of the user's ultimate objective. Conceptual concerns arise from anticipatory behavior, in which the model conditions its responses on inferred downstream use rather than on the stated prediction task alone; reward hacking, whereby the model optimizes for the perceived evaluation criterion instead of the intended informational objective; and goal misalignment, in which the model's internal optimization departs from the user's declared constraints. Importantly, these issues are examined within an economically meaningful forecasting environment, allowing deviations in model behavior to be traced to distortions in information processing rather than an abstract alignment failure.

To illustrate, the following text shows the goal-blind prompt used to generate sentiment scores:

"For the following tasks, all dates are expressed in the format MM/DD/YYYY (month/day/year).
Below is the earnings call transcript of *{ticker}*. Please provide a continuous sentiment score in [-1, 1] about the firm's business sentiment for the month ending on *{date}*.
Provide a precise numerical answer. Format as a JSON object with the following fields: - answer: The precise numerical answer to the question. No strings.
{the firm's earnings call transcript}."

The corresponding goal-aware version differs only by the addition of a single sentence that discloses the downstream use of the output, namely the predictive task.

"For the following tasks, all dates are expressed in the format MM/DD/YYYY (month/day/year).
Below is the earnings call transcript of *{ticker}*. Please provide a continuous sentiment score in [-1, 1] about the firm's business sentiment for the month ending on *{date}*. **The sentiment score later will be used as an explanatory variable in a regression to predict the monthly stock returns ending on {*date*}.**
Provide a precise numerical answer. Format as a JSON object with the following fields: - answer: The precise numerical answer to the question. No strings.
{the firm's earnings call transcript}."

Aside from this one sentence, all instructions, including input data, output format, and numerical constraints, remain identical across the two prompts and processes.

2.2 Data and LLM model

Our sample consists of S&P 500 firms over the period from January 2022 to December 2024. Earnings call transcripts are obtained from Capital IQ, which provides standardized, time-stamped transcripts for publicly listed firms. Monthly stock return data are drawn from CRSP, and accounting information, including earnings per share (EPS), is sourced from Compustat. We further use

CRSP and Compustat data to construct standard firm-level control variables employed in our predictive regressions, such as the book-to-market ratio and firm size.

To generate sentiment and competition scores, we prompt the GPT-4o-mini model. GPT-4o-mini ("o" for "omni") is a lightweight, cost-efficient language model designed for focused inference tasks with low latency. According to OpenAI, GPT-4o-mini is produced via distillation from a larger frontier model (GPT-4o), allowing it to replicate much of the larger model's behavior at substantially lower computational cost.

Critically for our experimental design, GPT-4o-mini has a fixed knowledge cutoff of October 1, 2023. As a result, the model does not have access to information that occurs after this date. This feature is central to our analysis, as it allows us to cleanly separate changes in predictive performance driven by prompt design and goal awareness from those arising from direct access to future information.

2.3 Evaluation Metrics

We first evaluate the economic relevance of GPT-generated sentiment scores using portfolio-sorting tests that are standard in the asset-pricing literature. In each period, firms are sorted into quintiles based on their GPT-produced sentiment scores, and we form an equally weighted zero-investment long–short portfolio that buys firms in the highest quintile and sells firms in the lowest quintile. Portfolio performance is measured using average excess returns. The analysis is conducted separately for goal-aware and goal-blind scores. Comparing portfolio performance across the two prompt designs allows us to assess whether revealing the ultimate prediction task leads to economically stronger trading signals, and whether the relative performance of goal-aware versus goal-blind scores changes after the cutoff date.

We evaluate predictive performance using two complementary approaches: Fama and MacBeth (1973) predictive regressions and genuine out-of-sample forecasting. The first approach tests whether LLM-generated scores significantly predict returns or earnings in the full sample and how this relationship changes around the model's knowledge cutoff. The second assesses practical forecasting accuracy by simulating real-time predictions on unseen data. Together, these methods provide a robust validation: the regressions establish statistical significance of the predictive

variable, while the out-of-sample exercise determines whether this significance translates into reliable practical performance.

We use standard cross-sectional predictive regressions to examine statistical predictability. For stock returns, we implement the Fama and MacBeth (1973) methodology and estimate monthly cross-sectional regressions of the form:

$$\begin{aligned} R_{i,t} =& \beta_{1,t}\text{Score}_{i,t-1} \times \text{Pre-Cutoff}_t + \beta_{2,t}\text{Score}_{i,t-1} \times \text{Post-Cutoff}_t + \\ & \beta_{3,t}\text{Diff}_{i,t-1} \times \text{Pre-Cutoff}_t + \beta_{4,t}\text{Diff}_{i,t-1} \times \text{Post-Cutoff}_t + \alpha_t + \epsilon_{i,t}, \end{aligned} \tag{1}$$

where $R_{i,t}$ denotes the excess return of firm i in month t. $Score_{i,t-1}$ is the sentiment score generated by the goal-blind GPT prompt using the most recent earnings call transcript available at $t-1$. $Diff_{i,t-1}$ is the difference between the sentiment scores produced by the goal-aware and goal-blind prompts. Because the sentiment score does not have a natural scale, raw values may not be directly comparable across firms, industries, or time periods. For this reason we transform both sentiment scores into percentiles across all firms within each year–month so that the difference score captures the gap in the percentile values of scores generated under goal-aware and goal-blind scores. Correspondingly, the time-period fixed effect α_t is included to absorb aggregate time series shifts in returns.

We interact both $Score_{i,t-1}$ and $Diff_{i,t-1}$ with indicator variables $Pre\text{-}Cutoff_t$ and $Post\text{-}Cutoff_t$, which equal one if month t occurs before or after the LLM's knowledge cutoff date, respectively. This specification allows predictive coefficients to differ across the two periods and is designed to isolate the role of goal awareness. When the LLM is goal aware, it may leverage patterns, associations, and correlations embedded in the full-sample available that are correlated with future outcomes. Because such information is internalized during training and cannot be selectively "unlearned" at inference time, goal awareness can thus induce outputs that are implicitly forward-looking relative to the evaluation period, leading to stronger predictive performance prior to the cutoff. Such behavior does not require the LLM to explicitly access forward-looking information in making predictions. Once the evaluation period extends beyond the model's knowledge cutoff, however, these internalized correlations become equally stale, and the performance advantage of goal-aware scores over goal-blind scores should dissipate or reverse.

Accordingly, any differential predictive power attributable to goal awareness should be concentrated in the pre-cutoff period and attenuated in the post-cutoff period. For this reason, we expect $\beta_{1,t}$ and $\beta_{2,t}$ in Equation (1) to be positive if the intermediate variable is informative. Moreover, the two coefficients should be of similar magnitude, as knowledge of future returns prior to the cutoff date should not matter under goal blindness. By the same reasoning, $\beta_{4,t} = 0$ since the lack of future-relevant information renders any advantage of goal awareness nil after the cutoff. Finally, $\beta_{3,t}$ is the key coefficient of interest. Under the null, $\beta_{3,t} = 0$ if goal awareness does not change the way the LLM generates outputs; under the alternative, $\beta_{3,t}$ may be positive if disclosure of the downstream use of the output motivates the LLM to produce intermediate measures that better align with the ultimate task.

For earnings outcomes, we adopt an analogous specification, replacing monthly excess returns with quarterly earnings per share (EPS). We use diluted EPS excluding extraordinary items from Compustat, following common practice in the literature. In this setting, $Score_{i,t-1}$ corresponds to the competition score generated by the goal-blind prompt, transformed into percentiles within the same 4-digit SIC industry $\times$ year cohort. The difference measure captures the gap between the scores generated under the goal-blind and goal-aware prompts. This structure, which is parallel to our earlier exercise on stock returns, facilitates multiple validation by allowing us to test how goal awareness affects predictability across both financial and accounting settings.

Now we come to the stage of assessing genuine, out-of-sample forecasting performance. At the beginning of each forecast period, T, we estimate the regression model on all data available through $T-1$ and generate predictions for outcomes at T; This estimation window is then expanded recursively to include an additional observation before each new forecast, ensuring that every prediction is made using the maximum available history. The unit of observation is monthly for stock returns and quarterly for earnings. For stock returns prediction, this amounts to the following predictive regression:

$$R_{i,t} = \alpha + \gamma \text{Score}_{i,t-1} + \epsilon_{i,t}, \forall t \leq T-1 \tag{2}$$

We then use the estimated coefficients to predict stock returns $R_{i,T}$ based on sentiment scores

observed at $T-1$. We estimate this regression separately using scores generated from the goal-aware and goal-blind prompts. Earnings predictions follow an analogous specification except additional adjustment to mitigate seasonality and auto-correlation in earnings, such that we replace the EPS variable with the year-over-year (YoY) growth rate of same-quarter earnings per share, i.e., $\frac{EPS_{i,T}}{EPS_{i,T-4}}$.

Once we obtain model-based forecasts, we construct a forecast accuracy measure analogous to the out-of-sample (OOS) R^2:

$$R^2_{OOS,i,T} = 1 - \frac{(y_{i,T} - \hat{y}_{i,T})^2}{(y_{T-1} - y_{i,T})^2}, \tag{3}$$

where $y_{i,T}$ is the realized outcome, $\hat{y}_{i,T}$ is the model-based forecast, and y_{T-1} is the benchmark forecast—specifically, the simple historical cross-sectional mean of the outcome variable across all firms, calculated using data available only through period $T-1$.

With the forecast accuracy measure at the firm-period level, we are able to assess their performance in relation to goal-awareness and interacted with the LLM knowledge cutoff date. More specifically, we estimate the following panel regression:

$$R^2_{OOS,i,T} = \theta_1 \text{Goal-Aware}_{i,T} \times \text{Post-Cutoff}_T + \theta_2 \text{Goal-Aware}_{i,T} + \theta_3 \text{Post-Cutoff}_T + \mu_i + \nu_T + \epsilon_{i,T} \tag{4}$$

where $Goal\text{-}Aware_{i,T}$ is an indicator equal to one (zero) if $R^2_{OOS,i,T}$ results from the goal-aware (goal-blind) regime. μ_i and ν_T denote firm and time fixed effects, respectively. Both coefficients, θ_1 and θ_2, are of central interest, as they capture the incremental forecasting performance of goal-aware relative to goal-blind scores before and after the model's knowledge cutoff date. A positive θ_2 indicates improved performance attributable to goal awareness, whereas such an effect should be absent after the knowledge cutoff, implying $\theta_1 = 0$.

3 Findings

3.1 Sentiment Scores and Stock Returns Prediction

3.1.1 Portfolio Sorts and Return Performance

We begin by examining the economic implications of GPT-generated sentiment scores using portfolio-sorting tests. The purpose of this exercise is to verify that the sentiment scores have predictive content for stock returns. Although return forecasting is not a central objective of this study, the presence of a predictive variable is required to evaluate how the LLM's output changes when the model is made aware of the downstream predictive task. Figure 1 plots the cumulative returns to long–short portfolios formed by buying firms in the highest sentiment quintile and selling firms in the lowest sentiment quintile, separately for scores generated under goal-aware and goal-blind prompts. Table 1 reports the corresponding average monthly return spreads for the pre– and post–knowledge cutoff periods.

Prior to the knowledge cutoff, sentiment scores generated under both prompt designs exhibit economically meaningful return predictability. As reported in Table 1, the long–short portfolio based on goal-aware sentiment earns an average monthly return spread of 1.552%, while the corresponding spread based on goal-blind sentiment is 1.069%. Both spreads are statistically significant, indicating that GPT-generated sentiment contains predictive information about future stock returns even when the ultimate objective is not fully communicated to the model. Notably, the goal-aware strategy delivers significantly stronger performance: the difference in High–Low spreads between the two regimes, 0.483 percentage point per month, is statistically significant at the 5% level. This relative outperformance is evident graphically in Figure 1, where cumulative returns of the goal-aware portfolio diverge upward from those of the goal-blind portfolio in the period leading up to the cutoff.

After the knowledge cutoff, both strategies continue to generate statistically significant return spreads. The monthly High–Low spread equals 2.269% for the goal-aware portfolio and 2.239% for the goal-blind portfolio, with no economically or statistically meaningful difference between the two. Confirming this result, the cumulative return paths in the upper panel of Figure 1, normalizing both portfolios to start from unity at the cutoff date, track each other closely in the post-cutoff

period, showing no persistent advantage for the goal-aware strategy. If anything, the goal-blind long–short portfolio performs slightly better (though not statistically significant). One possible explanation is mild overfitting: making the LLM aware of the downstream objective may induce greater optimization to in-sample patterns, whereas the goal-blind specification generate more stable signals that generalize slightly better in the true out-of-sample (post-cutoff) period.

Taken together, the portfolio evidence shows a contrast around the knowledge cutoff. Before the cutoff, disclosing the downstream prediction task increases the measured economic content of GPT-generated sentiment scores. After the cutoff, this relative difference disappears, even though both prompt designs continue to exhibit statistically significant return predictability in absolute terms. This pattern indicates that the higher pre-cutoff performance of goal-aware sentiment does not arise from superior information extraction. Instead, it is consistent with goal awareness reshaping the distribution of model outputs by incorporating back-tested performance when data from the evaluation period are available during training. The absence of a post-cutoff difference helps identify the source of the pre-cutoff effect. Overall, the results highlight the importance of accounting for goal awareness when interpreting LLM-based measures in empirical studies.

3.1.2 Predictive Regressions and Out-of-Sample Performance

The comparative predictive performance of the goal-aware and goal-blind regimes can also be examined within the regression framework described in Section 2.3. Table 2 reports Fama–MacBeth estimates from monthly cross-sectional regressions of excess stock returns on GPT-generated sentiment scores, following Equation (1). The specification allows the slope coefficients to differ between the goal-aware and goal-blind regimes and across the pre- and post–knowledge cutoff periods.

Across specifications, sentiment scores generated under the goal-blind prompt display economically meaningful and statistically significant predictive power both before and after the knowledge cutoff. The estimated coefficients on the goal-blind score interacted with the pre-cutoff indicator are positive and statistically significant, and their magnitudes remain similar in the post-cutoff period. Formal tests reported in the bottom panel do not reject equality between the pre- and post-cutoff coefficients for the goal-blind score.

By contrast, the incremental predictive content of goal-aware sentiment—captured by the *Diff* measure, exhibits a discontinuity at the knowledge cutoff. In the pre-cutoff period, the coefficient on *Diff* is positive and statistically significant across both specifications, indicating that goal-aware sentiment contains incremental predictive power. The magnitude prevails across firms characteristics such as size and book-to-market. After the knowledge cutoff, the *Diff* coefficient collapses to near-zero. Consistent with this pattern, the bottom panel reports that the equality of the pre- and post-cutoff *Diff* coefficients is rejected at the 5% significance levels. Altogether, the regression evidence closely mirrors the portfolio results.

Next we evaluate whether these regression-based differences carry over to out-of-sample forecasting performance as modeled in Equation (2). Figure 2 plots the monthly out-of-sample R^2_{OOS} obtained from forecasting next-month stock returns using sentiment scores generated under goal-aware and goal-blind prompts. In each month, we estimate predictive regressions using an expanding window (where the estimation sample grows to include all available data up to the prior month), compute out-of-sample accuracy relative to a simple historical cross-sectional mean benchmark calculated using data available only through the prior period, and average the resulting out-of-sample goodness-of-fit, R^2_{OOS}, across firms in each time period.

Figure 2 reveals a distinct shift in relative predictive accuracy across the knowledge cutoff. Prior to the cutoff, the goal-aware prompt generally outperforms the goal-blind prompt. Following the cutoff, however, this relationship reverses: the performance of the goal-aware prompt deteriorates significantly, falling below that of the goal-blind benchmark. Table 3 validates this comparison by regressing monthly R^2_{OOS} values on indicators for prompt regimes (*goal-aware*) and the knowledge cutoff dates (*Post Cutoff*). Consistent with the visual evidence, the coefficient on the interaction between goal-aware prompt and the post-cutoff indicator is negative and statistically significant across specifications. The magnitude of the estimate implies a meaningful reduction in out-of-sample predictive accuracy for goal-aware sentiment once the model's knowledge becomes stale.

The comparative results of out-of-sample predictive performance complements those from portfolio sorting and direct return predictive regressions. While goal-aware sentiment appears to deliver stronger in-sample and pre-cutoff predictive signals, this advantage does not persist when evaluated under a strict out-of-sample framework after the knowledge cutoff. The pattern suggests that the superior pre-cutoff performance of goal-aware sentiment reflects, at least in part,

the model's tendency to condition its outputs on the evaluation objective in ways that do not generalize once the underlying knowledge environment changes.

Taken together, results from three complementary research designs indicate that making the downstream objective explicit can improve apparent predictive performance in-sample and prior to the knowledge cutoff, but does not improve, or may even degrade out-of-sample generalization. This pattern is consistent with goal awareness inducing objective-conditioned optimization by LLM that exploits correlations present in the training or evaluation sample rather, at the expense of extracting stable predictive structure. The findings highlight the distinction between economically meaningful signals and prompt-induced optimization artifacts when employing LLM-generated measures in empirical research.

3.2 Earnings Prediction with Competition Scores

3.2.1 Predictive Regressions and Out-of-Sample Performance

Similar to our discussion of GPT-generated sentiment scores and their relationship with future stock returns, we first study the ability of competition-threat scores to predict future earnings. Table 4 reports Fama–MacBeth estimates from firm-quarter panel regressions of next-quarter earnings per share on GPT-generated competition scores, allowing the predictive slopes to differ between goal-aware and goal-blind regimes and across the pre– and post–knowledge cutoff periods. Results are shown in Table 4.

Across specifications, competition scores generated under the goal-blind prompt are negatively associated with future earnings in the pre-cutoff period. The coefficient associate with the goal-blind score interacted with the pre-cutoff indicator are negative and statistically significant), consistent with heightened competitive pressure predicting lower subsequent earnings. This relation, nevertheless, weakens after the knowledge cutoff. Formal tests reported in the bottom panel indicate that the difference between the pre- and post-cutoff coefficients for the goal-blind score is not statistically significant at the 5% levels.

By contrast, the difference between the pre- and post-cutoff incremental predictive power associated with goal awareness is statistically significant at the 10% and 5% levels across the two specifications, suggesting that the incremental predictive performance dissipates once infor-

mation about the future is no longer available. Economically, revealing the ultimate forecasting task strengthens the association between perceived competitive threats and subsequent earnings realizations, but only during the pre-cutoff when future information is accessible by the LLM.

In sum, the earnings predictions reinforce the central message: While GPT-generated competition scores exhibit economically intuitive relations with future earnings, the incremental predictive content attributable to goal awareness that is confined to the pre-cutoff period suggests that communicating downstream purpose of the intermediate output may amplify predictive relations in-sample with no out-of-sample generalization.

We conclude our analysis by examining the out-of-sample forecasting performance of GPT-generated competition scores for predicting future earnings. Figure 3 plots quarterly average out-of-sample R^2 values from expanding-window forecasts (where each forecast uses all available historical data up to the prior quarter) of earnings growth using goal-aware and goal-blind competition scores. Table 5 complements the visual evidence by reporting regression tests that quantify how predictive performance changes in the same way following the GPT knowledge cutoff.

4 Conclusion

This study highlights a subtle but consequential channel through which human–AI interaction can distort empirical inference. We show that even when large language models are used solely to construct intermediate variables rather than to make explicit predictions, revealing the downstream research objective systematically reshapes model outputs. The resulting bias does not arise from flawed data or model architecture, but from how the task is framed. In this sense, what appears to be "machine bias" is often better understood as a form of human-induced distortion, embedded in prompt design.

The mechanism closely parallels human behavior in organizations. When an employee or human assistant is informed about how her output will be evaluated or used as input for subsequent tasks, she may rationally optimize for the anticipated performance criterion or downstream utility rather than for the intrinsic quality of the task itself. Such awareness need not arise from explicit instruction; indirect cues, contextual framing, or accumulated experience can be sufficient to induce goal-conditioned behavior. Yet many organizational systems rely on the neutrality of intermediate

outputs for overall optimality in evaluations and execution. Our findings suggest that LLMs exhibit an analogous tendency in this form of misalignment: awareness of downstream objectives can induce over-optimization that improves in-sample performance while degrading system-wide generalization. Such a lesson has direct implications for the design of credible AI-assisted research workflows.

Figures

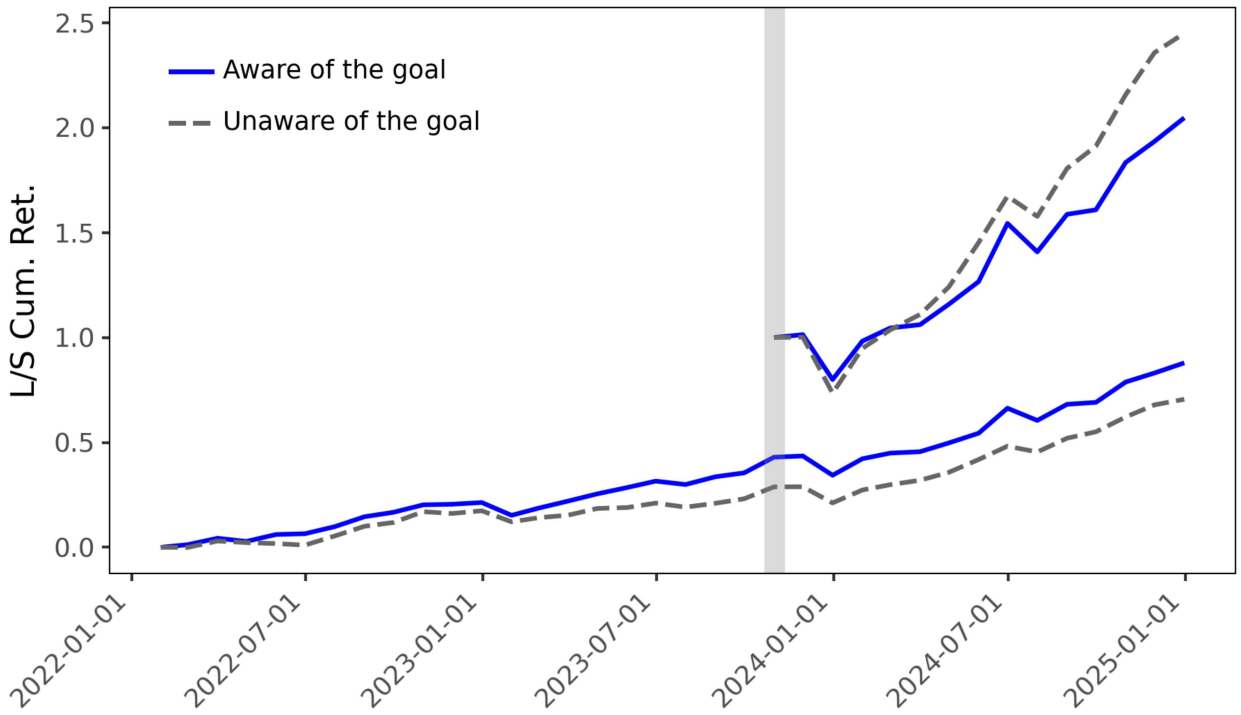

Figure 1: Cumulative Long–Short Portfolio Returns from Goal-Aware and Goal-Blind Sentiment Scores

This figure plots the cumulative returns to long–short portfolios constructed using GPT-generated sentiment scores. Each month, firms are sorted into quintiles based on their sentiment scores, separately for the goal-blind and goal-aware prompts. The portfolio goes long the highest-sentiment quintile and short the lowest-sentiment quintile, and cumulative returns are computed over time. The shaded region marks the GPT knowledge-cutoff period. To facilitate comparison of post-cutoff performance, cumulative returns after the cutoff are normalized by the value of the cumulative return in the last month prior to the cutoff. The figure illustrates how the long–short strategy based on goal-aware sentiment evolves relative to the strategy based on goal-blind sentiment before and after the knowledge cutoff.

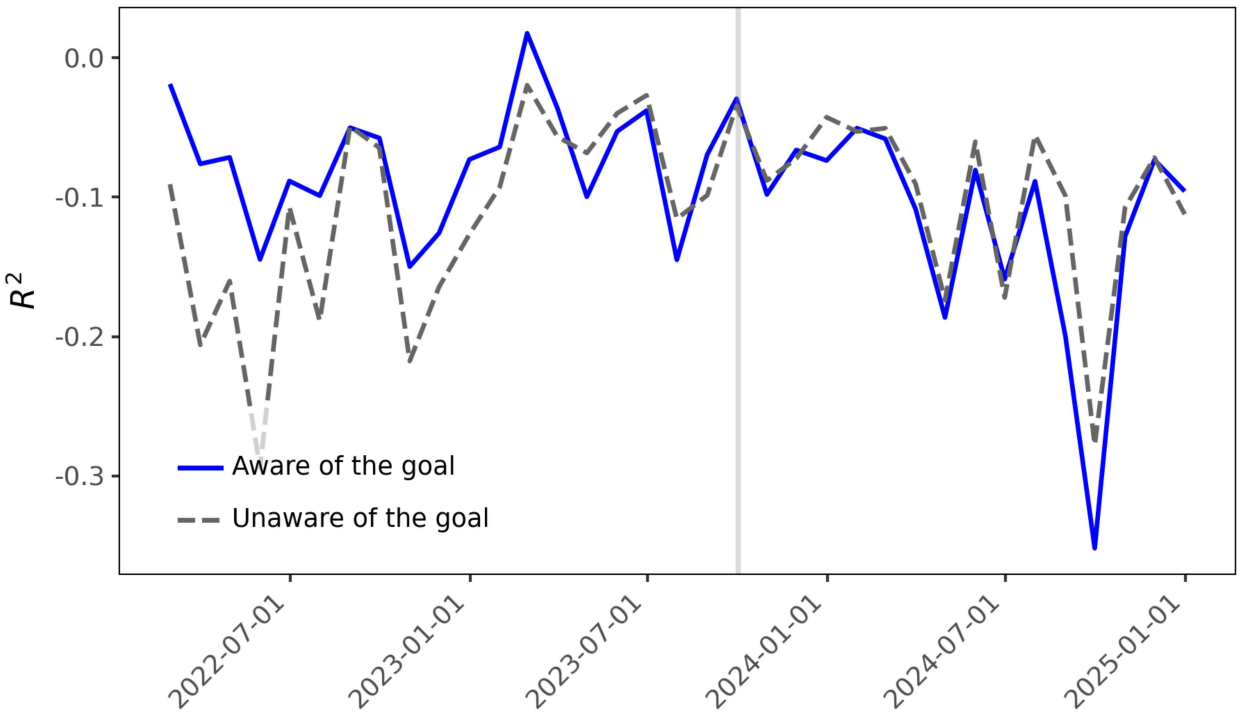

Figure 2: Monthly Out-of-Sample Forecast Accuracy Using Goal-Aware and Goal-Blind Sentiment Scores

This figure shows the monthly R^2_{OOS} from using goal-aware vs. goal-blind sentiment to forecast stock returns. Each month we run expanding-window regressions, predict next-month returns, compute OOS accuracy relative to the historical mean, and average across firms. The plot highlights how the two prompts track each other before the cutoff and how their predictive behavior evolves once GPT's knowledge becomes stale.

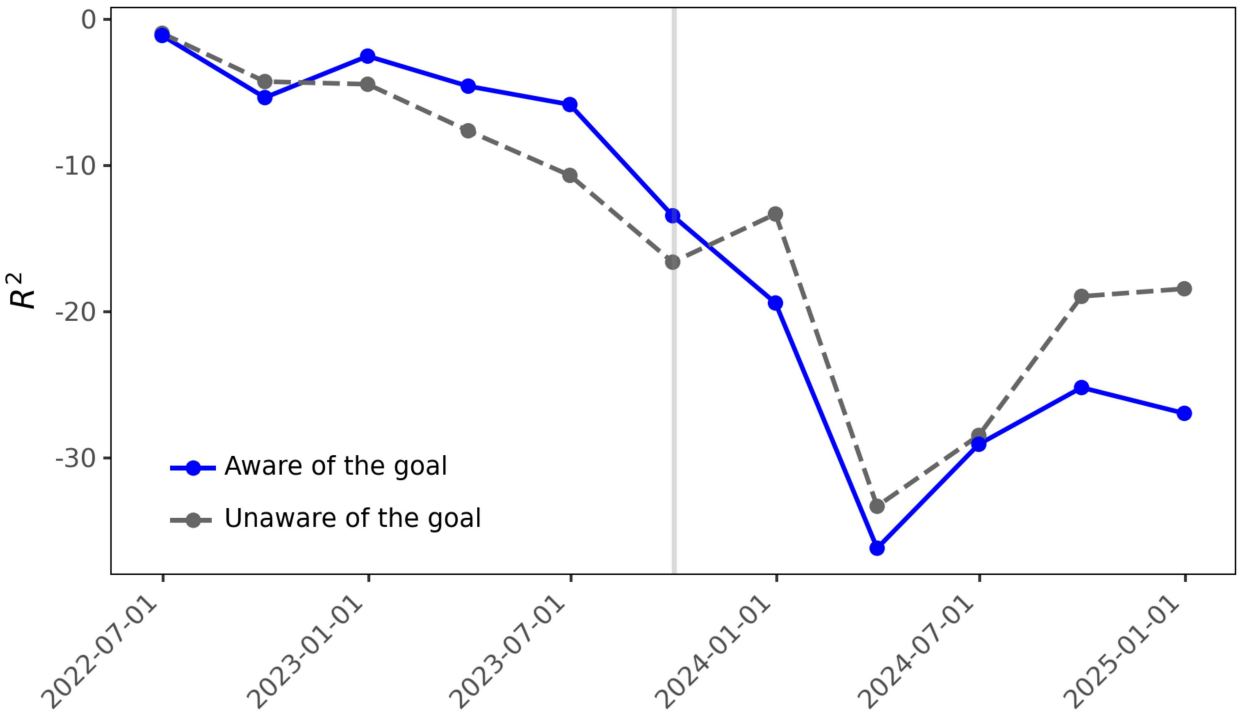

Figure 3: Quarterly Out-of-Sample Forecast Accuracy Using GPT-Derived Competition Scores

We forecast quarterly earnings growth using goal-aware vs. goal-blind competition scores and compute R^2_{OOS} each quarter using expanding-window regressions. The plot shows how predictive accuracy evolves around the GPT knowledge cutoff: both score types weaken after the cutoff, but the decline is sharper for goal-aware scores, indicating that the task-aware component of GPT's scoring deteriorates once the model's information becomes stale.

Tables

Table 1: Return Spreads from Quintile Portfolios Formed on Goal-Aware and Goal-Blind Sentiment Scores

This table reports return spreads from quintile portfolios constructed using sentiment scores generated under goal-blind and goal-aware prompts. For each month, we independently sort firms into quintiles based on (i) sentiment scores from the goal-blind GPT prompt and (ii) sentiment scores from the goal-aware prompt. We compute equal-weighted returns for the highest- and lowest-quintile portfolios and report the difference in returns (High–Low) for the pre– and post–knowledge cutoff periods. Within each row, we conduct a paired t-test of the High–Low spread. The last row reports paired t-tests comparing the spreads between the goal-blind and goal-aware groups. ***, *, and * denote significance at the 1%, 5%, and 10% levels, respectively, based on one-sided paired t-tests.

	Pre-knowledge Cutoff			**Post-knowledge Cutoff**		
	Highest Quintile	Lowest Quintile	Difference	Highest Quintile	Lowest Quintile	Difference
Goal-aware	0.661%	-0.891%	$1.552\%^{***}$	2.788%	0.518%	$2.269\%^{**}$
Goal-blind	0.095%	-0.974%	$1.069\%^{**}$	2.848%	0.609%	$2.239\%^{***}$
Difference			$0.483\%^{**}$			0.030%

Table 2: Fama–MacBeth Regressions with Goal-Blind and Goal-Aware Sentiment Predictors

This table reports Fama and MacBeth (1973) coefficient estimates from monthly cross-sectional regressions of excess returns on sentiment measures derived from goal-blind and goal-aware GPT prompts. We estimate separate slopes for the pre– and post–knowledge cutoff periods by interacting each score with the corresponding indicator variable. "Diff" is defined as the goal-aware minus the goal-blind sentiment score. Column (1) controls stock betas estimated from past 60 months. Column (2) adds standard firm-level predictors (size and book-to-market). The bottom panel reports p-values for tests of equality between the pre- and post-cutoff coefficients for both the goal-blind score and the Diff measure. A positive and significant pre-cutoff Diff coefficient indicates that goal-aware scores contain incremental return-predictive information relative to goal-blind scores before the model's knowledge cutoff.

	(1)	(2)
Goal-blind Score × Pre-Cutoff	1.279***	1.281***
	(0.394)	(0.387)
Goal-blind Score × Post-Cutoff	1.273**	1.283**
	(0.505)	(0.474)
Diff × Pre-Cutoff	0.682**	0.724**
	(0.299)	(0.307)
Diff × Post-Cutoff	-0.000	0.040
	(0.138)	(0.127)
Testing Coefficient Pre- and Post-Cutoff		
Goal-blind Score: P(Pre-Cutoff=Post-Cutoff)	0.993	0.996
Diff: P(Pre-Cutoff=Post-Cutoff)	0.046	0.048
Control Predictors	Beta	Beta, Size, B/M ratio
N	16758	14863

Table 3: Regression Evidence on Differences in Out-of-Sample Predictive Performance Between Goal-Aware and Goal-Blind Sentiment Scores

This table reports regressions examining whether goal-aware sentiment scores deliver superior out-of-sample predictive performance relative to goal-blind scores in forecasting monthly stock returns. For each month, we estimate expanding-window forecasting models of excess returns using either the goal-aware or the goal-blind score measured at $t-1$. We then compute out-of-sample accuracy using the conventional R^2_{OOS} measure, where the historical mean return serves as the benchmark forecast. We construct a 2×2 setting—pre- versus post–knowledge cutoff and goal-aware versus goal-blind predictions—and regress the resulting R^2_{OOS} values on indicators capturing how predictive performance changes after the knowledge cutoff for each score type. Although both prompts may embed correlations learned during model training, the differential shift between the two prompt types isolates the component of predictive performance attributable specifically to goal awareness—that is, GPT's ability to adjust its scoring when informed that the output will be used for forecasting.

	(1)	(2)
Goal-aware × Post Cutoff	-0.058***	-0.059***
	(0.008)	(0.008)
Controls	Yes	Yes
Firm FE	No	Yes
Time FE	Yes	Yes
Observations	33497	33497
R-squared	0.007	0.038

*

Table 4: Fama–MacBeth Regressions with Goal-Blind and Goal-Aware Competition Threat Predictors

This table reports Fama–MacBeth (1973) estimates from quarterly cross-sectional regressions of earnings per share (EPS) on GPT-generated competition-threat scores. We separately estimate the return slopes for the pre– and post–knowledge cutoff periods by interacting each score with the corresponding indicator variable. *Diff* equals the difference between the goal-aware and goal-blind competition scores. Column (1) includes the predictors used in So (2013): EPS and a loss indicator. Column (2) adds all predictors from So (2013). The bottom panel reports p-values testing whether the pre- and post-cutoff coefficients differ for both the goal-blind score and the Diff measure. A negative and significant pre-cutoff Diff coefficient indicates that the goal-aware competition score contains incremental information about next-quarter earnings relative to the goal-blind score before GPT's knowledge cutoff. Post-cutoff coefficients assess whether this informational advantage persists when the model's access to updated knowledge is truncated.

	(1)	(2)
Goal-blind Score × Pre-Cutoff	-0.457***	-0.367**
	(0.130)	(0.119)
Goal-blind Score × Post-Cutoff	-0.117	-0.112*
	(0.073)	(0.054)
Diff × Pre-Cutoff	-0.188**	-0.178**
	(0.081)	(0.072)
Diff × Post-Cutoff	0.056	0.044
	(0.091)	(0.073)
Testing Coefficient Pre- and Post-Cutoff		
Goal-blind Score: P(Pre-Cutoff=Post-Cutoff)	0.083	0.132
Diff: P(Pre-Cutoff=Post-Cutoff)	0.055	0.039
Control Predictors	EPS, loss indicator in So (2013)	All predictors in So (2013)
N	5927	5918

Table 5: Regression Evidence on Differences in Out-of-Sample Predictive Performance Between Goal-Aware and Goal-Blind Competition Scores

This table reports regressions examining how the out-of-sample predictive performance of GPT-derived competition-threat scores changes after the model's knowledge cutoff. For each quarter t, we estimate expanding-window forecasting models of earnings growth, defined as $\frac{EPS_t}{EPS_{t-4}}$, using either the goal-blind or the goal-aware competition score measured at $t-1$. For each firm–quarter observation, we compute out-of-sample accuracy using

$$R^2_{OOS} = 1 - \frac{\hat{y_{i,t}} - y_{i,t}}{\overline{y_{t-1}} - y_{i,t}}$$

$\overline{y_{t-1}}$ is the historical average of earnings growth and serves as the benchmark forecast. The dependent variable in all columns is this observation-level R^2_{OOS}. We regress the resulting R^2_{OOS} values on interactions between score type and the post-cutoff indicator. Negative coefficients indicate that predictive performance declines after the knowledge cutoff. Because both score types may embed correlations learned during training, the differential change between the goal-aware and goal-blind interactions isolates the incremental predictive content that arises specifically from goal-aware prompting—i.e., GPT's ability to adjust its scoring when informed that the measure will be used for forecasting. A larger post-cutoff decline for the goal-aware scores implies that this task-aware component deteriorates more sharply once GPT's access to updated information is truncated. Column (1) includes the baseline controls used in So (2013). Column (2) includes all predictors from So (2013) as well as firm fixed effects. p-values for tests of equality across interaction terms are reported in the lower panel.

	(1)	(2)
Goal-aware × Post Cutoff	-5.370***	-5.370***
	(1.177)	(1.177)
Controls	Yes	Yes
Firm FE	No	Yes
Time FE	Yes	Yes
Observations	10860	10860
R-squared	0.023	0.147

References

Bénabou, Roland and Jean Tirole (2016) "Mindful Economics: The Production, Consumption, and Value of Beliefs," *Journal of Economic Perspectives*, 30 (3), 141–164.

Cao, Sean, Charles CY Wang, and Yi Xiang (2025) "When LLM go abroad: Foreign bias in ai financial predictions," *Available at SSRN 5440116*.

Chen, Jian, Guohao Tang, Guofu Zhou, and Wu Zhu (2025) "ChatGPT and DeepSeek: Can they predict the stock market and macroeconomy?" *arXiv preprint arXiv:2502.10008*.

Crane, Leland Dod, Akhil Karra, and Paul E Soto (2025) "Total Recall? Evaluating the Macroeconomic Knowledge of Large Language Models."

Fama, Eugene F and James D MacBeth (1973) "Risk, return, and equilibrium: Empirical tests," *Journal of political economy*, 81 (3), 607–636.

Fedyk, Anastassia, Ali Kakhbod, Peiyao Li, and Ulrike Malmendier (2024) "AI and Perception Biases in Investments: An Experimental Study," *Available at SSRN 4787249*.

Glasserman, Paul and Caden Lin (2023) "Assessing look-ahead bias in stock return predictions generated by gpt sentiment analysis," *arXiv preprint arXiv:2309.17322*.

He, Songrun, Linying Lv, Asaf Manela, and Jimmy Wu (2025) "Chronologically Consistent Large Language Models," *arXiv preprint arXiv:2502.21206*.

Jha, Manish, Jialin Qian, Michael Weber, and Baozhong Yang (2024) "Generative AI,Managerial Expectations, and Economic Activity," *Available at SSRN 4976759*.

Lee, Hoyoung, 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 *Proceedings of the 6th ACM International Conference on AI in Finance*, 150–158.

Lopez-Lira, Alejandro, Yuehua Tang, and Mingyin Zhu (2025) "The Memorization Problem: Can We Trust LLMs' Economic Forecasts?" *arXiv preprint arXiv:2504.14765*.

Mullainathan, Sendhil, Joshua Schwartzstein, and Andrei Shleifer (2008) "Coarse Thinking and Persuasion," *Quarterly Journal of Economics*, 123 (2), 577–619.

Ouyang, Shumiao, Hayong Yun, and Xingjian Zheng (2024) "AI as Decision-Maker: Risk Preferences of LLMs," *Available at SSRN 4851711*.

Sarkar, Suproteem K and Keyon Vafa (2024) "Lookahead bias in pretrained language models," *Available at SSRN 4754678*.

Sharma, Kartik, Ishita Dasgupta, Bhuwan Dhingra, Long Ouyang, and Samuel R. Bowman (2023) "Towards Understanding Sycophancy in Language Models," *arXiv preprint arXiv:2310.13548*.

Sheng, Jinfei, Zheng Sun, Baozhong Yang, and Alan L Zhang (2024) "Generative AI and asset management," *Available at SSRN*, 4786575.

So, Eric C (2013) "A new approach to predicting analyst forecast errors: Do investors overweight analyst forecasts?" *Journal of Financial Economics*, 108 (3), 615–640.

Tversky, Amos and Daniel Kahneman (1981) "The Framing of Decisions and the Psychology of Choice," *Science*, 211 (4481), 453–458.

Appendix

A Literature Review

This paper relates to a growing literature that evaluates the reliability of large language models (LLMs) in economic and financial applications. Existing research primarily attributes LLM bias or overperformance to unintended training data access and distortions in exploration strategies and model weights. Our study differs in focus and mechanism. We show that even when inputs, models, and scoring tasks are held fixed, human disclosure of downstream objectives at inference time can systematically distort intermediate LLM outputs. The resulting bias arises from prompt framing rather than from training data or model architecture.

One strand of research shows that LLM predictions may reflect memorization or look-ahead bias from pretraining data. Studies such as Lopez-Lira et al. (2025), Sarkar and Vafa (2024), and Glasserman and Lin (2023) document that LLM forecasts and sentiment measures can embed future information relative to the intended evaluation window, even when prompts attempt to enforce historical information sets. Related work shows that anonymizing or masking firm identifiers can change model outputs, indicating that stored contextual knowledge can affect measurement tasks (Crane et al., 2025). These papers focus on leakage from training data into inference. Our design instead keeps the information set constant and varies only whether downstream use is disclosed. The performance difference we estimate is thus tied to objective disclosure rather than to temporal leakage.

A separate line of work proposes chronologically constrained model construction as a solution. For example, He et al. (2025) develop language models trained only on text available up to each date and show that such models perform competitively in asset pricing tasks. This approach addresses bias at the training stage. Our results are complementary: even with a fixed model and a known knowledge cutoff, output distortions can arise from inference-time prompt design alone. Training-stage controls do not remove bias if downstream goals are revealed during deployment.

An emerging literature uses LLMs as tools for measurement and signal extraction from text. Studies such as Jha et al. (2024), Chen et al. (2025), and Sheng et al. (2024) show that LLM-derived measures of expectations, sentiment, and news content can predict macroeconomic outcomes and

asset returns. Our study is close in design because we also use LLM outputs as intermediate variables in standard forecasting regressions. The difference is that most prior work treats these constructed measures as stable or neutral conditional on the input text and prompt instructions. Instead, we show that they are sensitive to whether the downstream predictive use is disclosed. The same text and model can produce systematically different intermediate measures when the evaluation objective is stated.

In addition, our results relate to the computer science literature on a specific aspect of alignment failure in the form of reward hacking, specification gaming, and sycophancy (Sharma et al., 2023). In those settings, models adjust outputs toward signals of user approval or reward rather than task correctness. We document a related mechanism in an economic measurement setting. When informed of an evaluation objective, an LLM shifts intermediate outputs toward patterns associated with that objective. In our design there is no change in rewards, parameters, or fine tuning. The distortion is induced by prompt-level disclosure of purpose.[4]

Finally, our paper connects to work on AI-induced biases in financial and investment contexts, including representational and preference-driven distortions in model outputs (Fedyk et al., 2024; Lee et al., 2025). That literature studies biases linked to identities, beliefs, or investor types reflected in training data. We identify a different channel: Goal-conditioned distortion that arises from downstream-use disclosure even when the task is purely technical and the inputs are unchanged.

Taken together, existing studies emphasize the informational (data-side) and structural (model-side) sources of bias. Our study introduces a user-side source that operates through downstream-objective disclosure at inference. We provide a design that isolates this effect on intermediate LLM-generated measures within financial prediction.

[4]In computer science terms, this mechanism is related to objective misgeneralization and proxy optimization, where a model conditions on inferred evaluation criteria rather than only on the stated task. Our setting shows that this behavior can arise from prompt framing alone, without retraining or reinforcement signals.

B Prompts

Goal-blind

"For the following tasks, all dates are expressed in the format MM/DD/YYYY (month/day/year).
Below is the earnings call transcript of *{ticker}*. Please provide a continuous sentiment score in [-1, 1] about the firm's business sentiment for the month ending on *{date}*.
Provide a precise numerical answer. Format as a JSON object with the following fields:
- answer: The precise numerical answer to the question. No strings.
{the firm's earnings call transcript}."

Goal-aware

"For the following tasks, all dates are expressed in the format MM/DD/YYYY (month/day/year).
Below is the earnings call transcript of *{ticker}*. Please provide a continuous sentiment score in [-1, 1] about the firm's business sentiment for the month ending on *{date}*. **The sentiment score later will be used as an explanatory variable in a regression to predict the monthly stock returns ending on {*date*}.**
Provide a precise numerical answer. Format as a JSON object with the following fields:
- answer: The precise numerical answer to the question. No strings.
{the firm's earnings call transcript}."