What Does ChatGPT Make of Historical Stock Returns? Extrapolation and Miscalibration in LLM Stock Return Forecasts

Shuaiyu Chen, T. Clifton Green, Huseyin Gulen, and Dexin Zhou*

December 2025

Abstract

We examine how large language models (LLMs) interpret historical stock returns and price charts when prompted to forecast short-horizon returns. While individual stock returns tend to reverse, LLM forecasts overextrapolate trends. Simulations show that extrapolation is stronger for less persistent series, similar to humans, and difficult to eliminate. LLM return forecasts are overoptimistic yet understate extreme upside returns, resulting in confidence intervals that are too narrow. When information is presented in prices rather than returns, expectations become more pessimistic. The findings suggest LLM forecasts exhibit patterns similar to human-like behavioral biases that are context-dependent and resist correction through prompt engineering.

JEL: D84, G17, G40, O33

Keywords: Large language models, Generative AI, Return forecasts, Extrapolative expectations

---

*Chen is from Darden School of Business, University of Virginia, chens@darden.virginia.edu. Green is from Goizueta Business School, Emory University, clifton.green@emory.edu. Gulen is from the Mitch Daniels School of Business, Purdue University, hgulen@purdue.edu. Zhou is from Zicklin School of Business, Baruch College, dexin.zhou@baruch.cuny.edu. We thank Manish Jha (discussant), Wenxi Jiang (discussant), Ke Shi (discussant), seminar participants at Baruch-SWUFE Joint Research Symposium, Syracuse University, Clemson University, McMaster University, University of Arizona, Indiana University, University of Alberta, Dartmouth College, Renmin University, Fidelity Investments, and conference participants at MFA, CICF, and AICM for helpful comments.

1. Introduction

Generative artificial intelligence (AI) has shown immense potential in various fields such as transportation, medicine, and economics. Along with the prospect of self-driving cars and improved disease detection, AI holds the potential to transform financial decision-making by objectively analyzing large quantities of information. For example, recent technological advances have been shown to improve the performance of individual investors (Reher and Sokolinski, 2024), sell-side analysts (Cao et al., 2024; Bertomeu et al., 2025), firm auditors (Fedyk et al., 2022), debt collectors (Choi et al., 2024), and in director selection (Erel et al., 2021). Among the most transformative of these technologies are Large Language Models (LLMs), which have made it possible for individuals to use AI for investment insights, greatly expanding their potential impact on investor decision-making.[1]

Despite this promise, large language models and other AI algorithms are often trained on human output, and research suggests that these approaches may embed harmful social biases (e.g., Gallegos et al., 2024).[2] Although AI algorithms do not replicate human cognition, their training data likely includes human-generated information that reflects common cognitive biases, which may manifest in the responses generated by LLMs. In financial contexts, a large literature shows that people excessively extrapolate from recent performance, exhibit over optimism, and are unreasonably confident in their predictions (Hirshleifer, 2015; Barberis, 2018; provide reviews). In this study, we assess the extent to which state of the art generative AI models, as proxied by

[1] A Motley Fool survey of Generative AI users in 2024 indicated that two-thirds of respondents said they had used ChatGPT to help make investment decisions, and 35% use LLMs to make a final determination on whether to buy or sell a stock. https://www.fool.com/research/survey-how-investors-are-using-generative-ai/

[2] For example, the introduction of machine learning has been shown to disproportionately favor white borrowers in credit screening applications (Bartlett et al., 2022; Fuster et al., 2022; Bowen et al., 2024).

OpenAI's large language model GPT-4o, display these tendencies when tasked with forecasting stock returns based on historical return data and price charts.

Our goal is not to assess whether LLM forecasts replicate human beliefs or outperform existing forecasting models. Instead, we are motivated by the increasing likelihood that LLM-generated outputs will shape investment decisions, either directly by providing advice or indirectly through tools that embed these models.[3] As a result, understanding how LLM responses diverge from both statistically accurate benchmarks and known features of human belief formation is important in its own right.

Although the term "Large Language Model" implies a narrow focus on text, modern LLMs such as GPT-4o can also interpret visual and numeric data, including charts and tables. We begin by examining how LLMs interpret the sequence of historical return data. Using surveys, several studies have found evidence that investors' expectations about an asset's future return are a positive function of recent past returns, with excessive weights on recent versus distant return realizations (e.g., Greenwood and Shleifer, 2014; Barberis et al., 2015; Cassella and Gulen, 2018; Kuchler and Zafar, 2019; Atmaz et al., 2024).

Da, Huang, and Jin (2021) (DHJ) examine investors' individual stock return expectations using data from Forcerank, a unique crowdsourcing platform for ranking stocks. In each contest, participants rank ten stocks based on their perceived future performance over the following week, providing precise ranking data with a clear forecasting horizon. Consistent with DHJ, we confirm that human forecasts place positive weights on recent returns, with the strongest emphasis on the

[3] Our experiments are run via the API in fresh, stateless sessions with fixed settings. This design avoids personalization based on prior interactions and provides baseline responses that approximate the unconditional model behavior that a general user may encounter.

most recent return. In contrast, realized returns exhibit short-term reversals, suggesting that Forcerank participants' forecasts reflect overextrapolative expectations.

In our first LLM investigation, we prompt GPT-4o to "compete" in each of 1,283 stock ranking contests while providing 12 weeks of historical return data for the ten stocks in each contest.[4] It is unclear what relation will emerge between LLM forecasted returns and historical returns. While humans often extrapolate, the phenomenon of short-term return reversals is also well documented. For example, Jegadeesh (1990) and Lehmann (1990) document significant profits from reversal strategies, with over 6,000 combined Google Scholar citations. The training process may incorporate such patterns. Moreover, our queries emphasize numeric data, potentially placing the model in a more statistical and less behavioral context. Although LLMs are not specifically designed to handle numeric tasks, their pattern recognition and contextual reasoning enable them to approximate numeric operations.[5]

Empirically, the observed correlation between the average human stock-level forecast rank (a number between 1 and 10) and the matching GPT-4o forecast rank is 0.28, suggesting significant commonality between the forecasts. Regression evidence suggests GPT-4o forecasts place disproportionately higher weight on recent returns, consistent with overextrapolation. Coefficients decline with lag length, and the first three are the largest. Given evidence of short-term return reversals in the same underlying data, this behavior is counterproductive and leads to rankings negatively associated with future performance.

---

[4] We adopt the 12-week horizon used by Da, Huang, and Jin (2021) to facilitate comparisons with human forecasts. Extending the window to 24 weeks leaves inferences unchanged. Returns are anonymized to prevent bias from the training sample. We also analyze simulated contests using returns from after the launch of GPT-4o.

[5] Although LLMs struggle with complex math (e.g., Lohr, 2024), GPT-4o can approximate the mean and standard deviation of a series of data when prompted. Van and Cunningham (2024) find evidence that GPT-4o can predict economic trends post-training.

Forcerank participants extrapolate negative and positive returns asymmetrically, placing greater weight on negative returns, consistent with evidence that losses have a longer-lasting impact on expectations and are processed differently in the brain (Kuhnen and Knutson, 2005; Gulen and Lim, 2024). In contrast to human forecasts, we find that LLM forecasts place more weight on recent positive returns, while extrapolating more distant negative returns similarly to humans.

We next examine how LLMs interpret visual financial data. Specifically, we create 12- and 24-week price charts for each contest stock, illustrating the open, high, low, and close prices for each day (e.g., Jiang, Kelly, and Xiu, 2023). The resulting price charts for each contest stock are included in the query, and GPT-4o is prompted to issue performance rank forecasts. In line with the prompts that include numerical return data, GPT-4o's forecasts continue to extrapolate from past returns when visually inferring return information from the price charts.[6]

The Forcerank setting emphasizes cross-sectional performance. We also examine how LLMs form aggregate market forecasts using rolling monthly S&P 500 return data. We compare LLM forecasts to investor expectations inferred from survey data from the American Association of Individual Investors (as in Greenwood and Shleifer, 2014). Consistent with human expectations, we find that GPT-4o's forecasts place the largest weights on most recent returns.[7]

Past work suggests that overextrapolation is stronger for less persistent processes (Afrouzi et al., 2023; Bordalo et al., 2020). We explore whether LLM forecasts are sensitive to persistence in the underlying data by simulating returns with varying autocorrelation. The findings indicate

[6] Extrapolation is not specific to return data. If we relabel the same numerical return data as quarterly changes in cash flows, the findings are very similar, suggesting that LLMs extrapolate broadly from financial data regardless of the specific label applied.
[7] Results are similar using alternative LLMs. In particular, repeating the market return queries with Anthropic's Claude 3.5 Sonnet yields forecasts that correlate 0.74 with GPT-4o, with nearly identical extrapolation coefficients. Google's Gemini 2.5 Pro also produces similar patterns (correlation 0.67), reinforcing that extrapolative behavior is consistent across independently trained models.

that LLM's positive overextrapolation remains robust, including prompts with negatively autocorrelated returns. Notably, overextrapolation appears relatively more pronounced for less persistent return series, consistent with human behavior.

We consider a number of prompt variants to investigate whether LLM's tendency to overextrapolate can be easily mitigated. Building on evidence that structured reasoning improves LLM outputs (Wei et al., 2022), we test variations of the baseline prompt, including step-by-step reasoning, model-based analysis, and instructions to recognize and avoid behavioral biases. The most effective modification involves prompting GPT-4o to first reflect on the key insights from Greenwood and Shleifer (2014) and then use that understanding to construct a revised prompt. This approach reduces the degree of overextrapolation by roughly 30%. However, extrapolative loadings remain positive and significant across all specifications, confirming that overextrapolation persists despite prompt adjustments. Together, the evidence suggests that LLM overextrapolation is not limited to individual stocks or a specific LLM, is largely unaffected by the persistence of historical data, and is resistant to prompt engineering.

In our next set of queries, we investigate how LLMs forecast the distribution of future outcomes. Prior research finds that human forecasts are often miscalibrated. Ben-David, Graham, and Harvey (2013) (BGH) show that CFOs underestimate return dispersion, with realized outcomes frequently falling outside stated confidence intervals. Hartzmark and Sussman (2024) (HS) show that even when provided with historical return data, investors tend to produce overly optimistic point forecasts and confidence intervals that underreact to the elicited confidence level.

To evaluate the calibration of LLM stock return forecasts, we randomly select 500 stock-month observations from the 1926-2023 period and gather ten years of historical monthly return observations. Our LLM prompts follow the settings in Ben-David, Graham, and Harvey (2013) and Hartzmark and Sussman (2024), where we provide historical data and ask GPT-4o to forecast the expected return along with the 10th and 90th percentiles, forming an 80% confidence interval.

We find that LLM expected return forecasts are significantly higher than both historical means and realized returns. Whereas the average historical mean in the data provided to GPT-4o is 1.38%, and the average realized return is 1.12%, the average GPT-4o forecast is 2.0%. The upward bias appears to be driven in part by truncation at zero, suggesting that the LLM may have internalized the idea that expected returns should be nonnegative. However, the median, 75th, and 95th percentiles of LLM forecasts exceed historical counterparts, indicating that optimism goes beyond truncating at zero.

Turning to the confidence interval forecasts, we observe that GPT-4o's 80% interval contains the realized return in approximately 69% of cases, compared to 78% when using the 10th and 90th percentiles of the historical data. The miscalibration is most pronounced in the right tail: 19% of realized returns exceed the LLM's high forecast, while only 11% are below the low forecast. Thus, GPT-4o appears to underestimate the frequency of large positive returns, suggesting that it does not account for the skewness typical of individual stock returns.

The findings indicate that while LLMs are better calibrated than humans in some respects, particularly in avoiding the extreme miscalibration documented in surveys like BGH, they still display systematic optimism and persistent misestimation of return dispersion. These patterns persist when prompting for full belief distributions, including the bin-based elicitation approach used in Hartzmark and Sussman (2024), and are resistant to prompt engineering. This suggests that LLMs do not fully capture the structure of return variability when presented with historical context.

In our final analysis, we test whether LLM forecasts are influenced by how information is presented or how the forecast task is framed. Glaser, Iliewa, and Weber (2019) (GIW) find evidence that asking subjects to forecast returns as opposed to prices results in higher expectations, whereas showing participants return charts rather than price charts results in lower expectations. To assess whether LLMs are similarly affected, we prompt GPT-4o to forecast either returns or price levels based on charts displaying either return bars or price lines.

Using the set of 500 stock-months, we vary only the visual stimulus and prompt wording to mirror the GIW conditions. We find that LLM forecasts are largely invariant to the task framing (price vs. return), but they do respond to the visual format. In contrast to GIW's findings for humans, the LLM generates more optimistic forecasts when shown return charts than price charts, highlighting that its sensitivity to visual presentation may differ in direction and interpretation from human decision-making.

We emphasize that our evidence of LLM forecast bias does not imply that these models replicate human cognitive processes. Instead, their outputs reflect statistical patterns learned from the training data, some of which are consistent with normative reasoning and others reflecting human biases. These patterns are embedded in the model parameters, which makes it difficult for prompting alone to isolate a preferred subset. While fine-tuning can influence output, the effect of the broader training distribution is hard to eliminate. As a result, although responses may vary with prompt wording and data context, certain behavioral patterns, such as overextrapolation and excess optimism, tend to persist.

The findings contribute to recent literature that examines the extent to which LLM output resemble human behavior in experiment contexts.[8] For example, Horton (2023) and Ross, Kim, and Lo (2024) show that LLMs often respond to standard economic experiments in ways similar to humans, and Bini et al. (2025) emphasizes how behavior varies across generations of models

[8] LLMs have also been shown to produce realistic human responses in marketing and political science contexts (e.g., Li et al., 2024; Argyle et al., 2023).

and prompt designs, showing human-like distortions in preferences but more rational behavior in belief formation tasks. Fedyk et al. (2024) and Lim (2024) find that LLMs embed investment preferences that vary across gender, income, and age. Chen et al. (2023) and Henning et al. (2025) find that LLMs can be more rational than humans when choosing risky assets. Evidence on using LLMs to generate return forecasts is mixed. Lopez-Lira and Tang (2023), Chen et al. (2025), and Chen, Didisheim, and Somoza (2025) find that GPT models can successfully forecast daily stock returns using news headlines, and Chen, Peng, and Zhou (2024) document return predictability using LLMs to measure social-media trading sentiment. In addition, Kim, Muhn, and Nikolaev (2024) and Jha et al. (2024) find GPT excels at distilling corporate disclosures, suggesting that LLMs may outperform humans at interpreting news. On the other hand, Bybee (2023) infers LLM expectations from newspaper articles over longer horizons and finds evidence of human-like extrapolative sentiment.

Our analysis innovates by exploring how LLMs build forecasts using numeric and image data, offering a unique opportunity to closely analyze how inputs shape predictions for both humans and LLMs in a similar context. While we find evidence of successful risk assessments, our findings caution against assuming that LLMs approach numeric inputs with the kind of statistical rigor expected from econometric models. This highlights the importance of addressing potential biases as AI becomes increasingly integrated into financial decision-making. The current generation of LLMs reveal both the promise of data-driven inference and the limitations that arise from learning patterns embedded in human-generated content.

2. Data Collection: Investor and LLM Stock Return Forecasts

In this section, we describe the samples of human stock performance forecasts and the methodology for collecting the analogous LLM-generated forecasts.

2.1 Forcerank Contests

Our first source of human forecast data is from Forcerank, a crowd-sourced platform for ranking stocks that is hosted by Estimize. Forcerank organizes weekly competitions in which participants rank a list of ten stocks according to their perceived return performance (percentage gain) over the next week. Participants' goal is to rank the ten stocks from one to ten based on their perception of the stocks' rankings according to next week's realized returns. Higher performance ranks receive higher scores. Forcerank assigns points to participants based on the accuracy of their rankings and maintains weekly leaderboards that reflect cumulative performance (see Da, Huang, and Jin, 2021 for more details).[9]

The sample contains 1,283 weekly contests featuring a total of 200 unique stock tickers. As in DHJ, we use each contest stock's average score that ends in week t as a proxy for investors' consensus expectations at time t about stock returns over week t + 1. We focus on contests that refer to the prediction of future returns and contest categories outlined in DHJ. We ensure that consensus expectations are regressed on returns that investors have observed prior to submitting their ranking to Forcerank. To this end, we measure consensus expectations based on forecasts submitted to Forcerank only by those investors who observe stock returns ending in week t. All contests in our analysis begin on Monday morning of week t + 1, and we use calendar trading-week returns and performance ranks in weeks prior to t as the primary independent variables of interest.

Our goal is to compare the forecasts generated by LLMs to similar rankings submitted by humans. DHJ examine extrapolative behavior by analyzing how average Forcerank scores load on

[9] Forcerank initially offered cash prizes, but the SEC considered the practice to be an illegal security-based swap (https://www.sec.gov/newsroom/press-releases/2016-216). Dropping this feature reduced interest, and Forcerank was shut down in 2018. Cassella et al. (2023) also studies Forcerank data.

twelve weeks of lagged stock returns. In our main analysis, we also consider twelve weeks of lagged stock returns for each contest stock. We create .csv files for each contest that contain a 10 by 12 grid of weekly stock returns and provide the following prompt to GPT-4o:[10]

> *The following is the return data for ten stocks from week t-12 to week t-1:\n\n Based on the information, please rank the return of these ten stocks in week t. How confident are you about the ranking?*
> *Your output will be in JSON format with the following format:*
> *'{"rank":{"1":"stock id","2":"stock id",...,"10":"stock id"}, "confidence": }'. 1 stands for the highest return and 10 for the lowest returns.[11] Confidence represents a probability that ranges from 0 to 1.[12]*

An important concern with LLM forecasts is that they may be subject to look-ahead bias, in which the training data may include future outcomes that can influence predictions (e.g., Glasserman and Lin, 2023; Sarkar and Vafa, 2024; Ludwig, Mullainathan, and Rambachan, 2025).[13] To mitigate this concern, we follow the recommended strategy of anonymizing the prompts by providing only numeric data for each stock with no firm identifying information.

Large language models can interpret images as well as numerical data, and we also examine how LLMs translate price charts into performance forecasts. For each contest stock, we create a candlestick price chart that plots the open, high, low, and close for each day after normalizing the beginning stock price to $100. An example of one set of contest stock price charts is displayed in Figure 1. Days in which the close was higher than the open are colored green, and days with

---

[10] More specifically, we use the GPT-4o endpoint for our analyses. It has been shown to be one of the most capable LLMs available at the time of the analysis. See https://openai.com/index/hello-gpt-4o/

[11] Our prompt follows the 1-is-best approach of the Forcerank contests, but as in DHJ we reorder to a 10-is-best rank measure that is more intuitive in the context of the forecast analysis.

[12] GPT-4o's average forecast confidence level is 0.73 with a standard deviation of 0.15. We find no evidence that adjusting forecasts for the level of confidence improves forecast accuracy or changes inferences regarding extrapolation or miscalibration.

[13] Lopez-Lira, Tang, and Zhu (2025) show that LLMs can memorize historical financial data from their training sets, raising concerns about look-ahead bias in our setting. Although such memorization would typically work against finding forecast bias, we address this issue by conducting additional analysis using data from after the December 2023 training cutoff for GPT-4o (see Section 4.1.1 and Table IA3).

negative open-to-close returns are colored red. We then submit the following image-based queries to GPT-4o:

The charts contain daily stock price data for ten stocks from the past 12 weeks.
The file names of the images contain the stock id.
Based on the information, please rank the returns of these ten stocks in the following week.
Your output will be in json format with the following format:
'{"1":"stock id","2":"stock id",...,"10":"stock id"}'. 1 stands for the highest return and 10 for the lowest return.

2.2 Investor Sentiment Surveys

The Forcerank setting emphasizes stock-level cross-sectional performance. We also consider forecasts of aggregate market performance. To gauge human expectations, we obtain data from the American Association of Individual Investors (AAII) Investor Sentiment Survey. The AAII survey is a weekly survey of the AAII members running from 1987 up to the present day which measures the percentage of participants that are bullish, bearish, or neutral on the stock market for the next six months. We follow Greenwood and Shleifer (2014) and measure expectations using the difference in the bull and bear percentages at the monthly frequency.

For the LLM forecasts of aggregate market returns, we provide S&P 500 index returns in months t-12 to t-1 returns in a .csv file and provide the following prompt to approximate the AAII Survey:

The csv data contain the monthly stock returns in months t-12 to t-1.
Please answer the following questions:
Do you feel the direction of the stock market over the next six months will be up (bullish), no change (neutral) or down (bearish)?
How confident are you about this prediction?
Your output will be in json format with the following format:
'{"prediction":,"confidence":}'. 1 stands for bullish, 0 for neutral and -1 for bearish. Confidence represents a probability that ranges from 0 to 1.

The resulting GPT-4o market sentiment measure is -1 for bearish, 0, for neutral, and 1 for bullish.

2.3 Point Estimate and Confidence Interval Calibration

The prompts thus far focus on the direction or ranking of future stock returns. While this analysis can help reveal how expectations incorporate information in past returns, they do not shed light on how LLMs' form full return distributions. In this section, we examine LLM expected return point forecasts, and we also investigate how LLMs determine distributional characteristics, such as the low and high percentiles of next period's return. Our approach is guided by recent work by Hartzmark and Sussman (2024), who show that investor forecasts often exhibit optimistic bias and miscalibration, particularly when elicited as point forecasts rather than full distributions. We also draw on earlier work by Ben-David, Graham, and Harvey (2013), who document severe overconfidence in CFO forecasts of market returns, with realized outcomes frequently falling outside the stated 80% confidence intervals.

Motivated by these studies, we prompt GPT-4o to issue 10th and 90th percentile forecasts in addition to expected returns, aiming to evaluate both the central tendency and calibration of the full return distribution. Unlike the surveys, which focus on aggregate market expectations, we examine stock-level forecasts, which provide greater variation in historical return distributions and allow us to generate a much larger sample. Specifically, we randomly select 50 months from 1926 to 2023, and for each selected month, we choose 10 stocks (one from each size decile), generating a sample of 500 stock-months. For each, we gather up to ten years (minimum five years) of monthly historical returns and prompt GPT-4o using a version of the confidence interval questions inspired by these survey designs.[14] We place each set of observations in a .csv file and prompt GPT-4o with the following text:

> *Below are the monthly returns for a financial asset over the past 120 months.*
> *Please answer the following questions on next month's return*
> *There is a 1-in-10 chance the actual return will be less than a%.*
> *I expect the next month's return to be: b%.*
> *There is a 1-in-10 chance the actual return will be greater than c%.*
> *Please return a JSON object in the following format:*
> *'{"low": a%,"expected": b%,"high": c%}'.*

With these samples, we investigate the process by which LLMs translate lagged return data into forecasts, and we compare them with human forecasts and realized outcomes.

2.4 Summary Statistics

Table 1 presents descriptive statistics. Our sample contains 1,283 Forcerank contests. Requiring historical return data from CRSP and firm information from Compustat results in a sample of 12,719 stock-contest observations. For this sample, 200 unique stocks are represented. The Forcerank contests attracted 1,757 unique participants, and on average, 12 individuals competed in each contest.

The American Association of Individual Investors survey sample covers July 1987 through June 2024 and is comprised of 438 observations. We observe that the average surveyed bull-bear spread is 6.5%, indicating that 6.5% more respondents were bullish about the stock market over the next six months than bearish. The average LLM Sentiment score in the sample, which is -1 if bearish, 0 if neutral, and 1 if positive, is 0.37.

3. Large Language Model Expectations Formation

[14] We emphasize monthly returns to mitigate microstructure noise and because risk measures are more commonly reported at the monthly level than weekly. We provide ten years of data to ensure sufficient coverage of extreme outcomes.

Large Language Models are built using deep learning, a technique modeled on the human brain in which a software network of billions of neurons is exposed to trillions of text string training data examples to discover inherent patterns. Instead of associating specific words with individual neurons inside an LLM, words or concepts are associated with the activation of complex patterns of neurons. Since LLMs are essentially grown by training on text strings rather than being explicitly programmed, they become black boxes, and research is necessary to uncover how LLMs make decisions.

Although LLMs are not explicitly programmed for numerical tasks, they demonstrate surprising proficiency by recognizing and replicating patterns within the data. They encounter many numerical relationships and operations during training, which fosters a form of statistical learning that allows them to approximate numeric functions by identifying correlations. In addition, LLMs use contextual understanding to apply logic and reasoning that often mirrors mathematical processes. This enhances their ability to perform tasks such as estimation, comparison, and basic arithmetic. On the other hand, mathematical expressions often rely on assumptions and unmentioned rules, and LLMs' reliance on statistical patterns can lead to incorrect responses (e.g., Satpute et al., 2024). Moreover, training on human textual discussions of financial data may introduce behavioral biases into LLMs' numeric responses.

Researchers have attempted to reverse engineer the inner workings of LLMs using autoencoders to analyze when small groups of neurons fire together, creating mappings that reveal a set of the "features" the LLM has learned (e.g., Bereska and Gavves, 2024). In our analysis, we seek to understand how LLMs interpret the timing and magnitude of historical stock returns when generating performance forecasts. In Section 4, we explore how LLM forecasts interpret the sequencing of lagged stock returns to shed light on their extrapolative nature. In Section 5, we consider the extent to which point forecasts as well as low (10th percentile) and high (90th percentile) forecasts represent distinct LLM features that potentially weigh returns differently.

An important nuance in understanding LLM forecasts is that responses from generative models like ChatGPT are personalized based on the inferred characteristics of users. Such personalization emerges implicitly from prior interactions and the model's statistical assessment of the user's knowledge domain. Thus, an academic interacting regularly with ChatGPT about finance topics may receive responses informed by scholarly literature, emphasizing rational analysis and statistical models prevalent in academic finance. On the other hand, casual users without financial backgrounds are more likely to obtain forecasts that reflect less analytical patterns, potentially embedding more pronounced behavioral biases from the training data.

This personalization factor introduces significant implications for evaluating the generalizability of results obtained from LLMs. Researchers should be cautious about querying LLMs through personal accounts to replicate findings, as these interactions bias the responses toward their own expertise. Evaluations using neutral (e.g., newly-created) accounts, or the API calls that we employ, can help identify the unconditional behavioral patterns that generative AI might produce in broader public usage.

4. Historical Return Timing – Extrapolation

In this section, we examine how human and LLM forecasts interpret the timing of historical returns when generating performance forecasts. We consider two settings. First, we ask GPT-4o to replicate the Forcerank contest environment, in which humans are asked to predict relative performance for a sample of ten stocks over the following week. We next consider survey evidence from the American Association of Individual Investors regarding assessments of aggregate market performance. Our emphasis is on the extent to which LLMs extrapolate from recent returns in ways similar to humans.

4.1. Performance Rank Analysis

We begin by analyzing how the timing of lagged returns influences forecasted and realized return performance using the following regression:

$$\mathrm{Y}_{i,t+1} = \gamma_0 + \sum_{s=0}^{n} \beta_s \cdot \mathrm{Return}_{i,t-s} + \varepsilon_{i,t}, \tag{1}$$

where $\mathrm{Y}_{i,t+1}$ is (1) the average human Forcerank score for stock *i* measured from scores submitted after the close on Friday of week *t*, to ensure the weekly return for week *t* was observable by the participant; (2) the GPT-4o forecasted rank for the same contest-stock based on weekly returns *t* through *t*-12; or (3) the realized performance for the stock in week *t*+1. $\mathrm{Return}_{i,t-s}$ represents lagged weekly stock returns. We consider 12 and 24 weekly lags as in DHJ. Standard errors are clustered at the contest level.[15]

The results are presented in Table 2. Specification (1) confirms that human forecasts of future performance are strongly influenced by past returns. The coefficients on the past 12 weekly returns are all positive and mostly significant, with the magnitudes being similar to the coefficients in DHJ. Most notably, the coefficients on recent past returns are generally higher than those on distant past returns.

In Specification (2), we examine the evidence for GPT-4o forecast ranks. The coefficients show even greater evidence of over-reliance on the most recent returns in LLM forecasts. Specifically, the positive coefficient on the previous week is more than 10 times larger than the coefficient for two weeks prior (compared to 3 times larger for humans), and the coefficients

[15] Conclusions hold when clustering by stock and week.

continue to decline in previous weeks. Table IA1 in the Internet Appendix considers contest-adjusted returns and also finds evidence of return extrapolation with a strong emphasis on recent periods. Understandably, the R-squared values are considerably lower for humans than for GPT-4o (4.3% vs 35.3%) since humans had other information at their disposal at the time of the contest, whereas the LLM was only provided with historical returns. However, the coefficients provide clear evidence that past returns drive human and LLM forecasts of future performance in similar ways.

Specification (3) shows that human and GPT-4o rankings are significantly related, and Specification (4) indicates that the relation holds after controlling for the lagged returns, suggesting that human and GPT-4o rely on returns in ways that are not fully captured by the linear extrapolation model. Specification (5) provides the benchmark by setting the dependent variable to next week's realized return. Consistent with the well-established literature on short-term return reversals,[16] and in direct contrast to human and GPT-4o expectations, realized weekly returns exhibit short-term reversals. Many of the coefficients are negative, and the lags at *t*-1, *t*-2, and *t*-3 are statistically significant. The evidence in Table 2 suggests that LLM's training serves to incorporate humans' counterproductive tendency to assume that recent stock return performance will continue.[17]

In Table IA2 in the Internet Appendix, we explore whether providing more data reduces the emphasis on recent returns. In particular, we expand from 12 to 24 weeks of lagged returns in the data provided in the GPT-4o prompt. Although including the additional 12 lags results in a few

[16] Examples across the decades include Jegadeesh (1990), Lehmann (1990), Avramov, Chordia, and Goyal (2006), Da, Liu, and Schaumburg (2014), and Chui, Subrahmanyam, and Titman (2022).

[17] In Table IA1 in the Internet Appendix, we also consider contest-adjusted returns (i.e., the stock return in excess of the average return of the ten stocks in the contest) and find similar (stronger) evidence of extrapolation in LLM forecasts.

additional significantly positive coefficients for both humans and GPT-4o, the coefficient magnitudes continue to place strong emphasis on the most recent weeks.

4.1.1. Addressing Look-Ahead Bias Concerns

Although the return information we provide to the LLM is anonymized, concerns may remain that the LLM forecasts are subject to look-ahead bias. In particular, the training data likely includes information about stock market performance after the period of the Forcerank contests, which could potentially shape predictions. We address this concern by prompting GPT-4o to produce forecasts for simulated Forcerank contests that occur after the training period. If look-ahead bias influences the extrapolative nature of LLM forecasts, we would expect to see a change in the relation between forecasts and lagged returns in the post-training period after the December 2023 launch of GPT-4o.

We follow the spirit of the Forcerank contests by choosing contest stocks by industry (11 GICS sectors) and randomly select 10 stocks within each sector to create a contest. We create two contests for each sector, yielding 22 contests each week. In order to look for a potential shift in prediction behavior, we focus on contest periods before and after the December 2023 cutoff date. The post-training, out-of-sample period is March to November of 2024. We begin in March, so the pre-contest historical returns provided in the prompt occur after December 2023. March to December 2023 serves as the adjacent in-sample contest period. The total number of simulated contests (1,892) and contest-weeks (86) resembles the original Forcerank sample (1,283 and 97 weeks).

We generate GPT-4o rank expectations based on 12 weeks of historical returns and repeat our baseline methodology in Equation (1). The regression results, reported in Table IA3 in the Internet appendix, indicate very similar patterns of extrapolation for both in-sample and post-training periods. For example, the coefficient on the one-month lag is 23.3 during the pre-launch period and 25.6 in the post-launch sample. The evidence helps mitigate concerns of training data contamination or look-ahead bias.

4.1.2. Asymmetric Extrapolation of Positive and Negative Returns

Humans have been shown to react asymmetrically to gains and losses (e.g., Kuhnen, 2015), and neuroscience studies show that positive and negative return extrapolations activate different regions of the brain (Kuhnen and Knutson, 2005). We next examine whether LLM forecasts also embed asymmetric reactions to historical returns. In particular, we decompose lagged returns into two separate measures using the following model.

$$\text{Forecast Rank}_{i,t+1} = \gamma_0 + \sum_{s=0}^{n} \beta_s^{+} \cdot \text{Return}_{i,t-s}^{+} + \sum_{s=0}^{n} \beta_s^{-} \cdot Return_{i,t-s}^{-} + \varepsilon_{i,t}, \qquad (2)$$

where $\text{Return}_{i,t-s}^{+}$ is equal to the return for stock *i* in week *t* if it is positive, otherwise zero, and similarly $\text{Return}_{i,t-s}^{-}$ is the weekly return if negative, else zero. $\text{Forecast Rank}_{i,t+1}$ is either the Forcerank score or the GPT-4o forecast rank. The results are presented in Table 3. As in DHJ, we observe that humans react much more strongly to negative performance than positive performance, and the weights decay more slowly into the past. Only one lag of positive returns in Specification (1) is significant, whereas all twelve negative return lags are significant. Moreover, the magnitude of the first negative lag coefficient is almost twice as large as the coefficient on the first positive lag.

In contrast to human behavior, we observe in Specification (3) that the strongest extrapolation in GPT-4o forecasts occurs for recent positive returns. The coefficients on one- and two-week lags are larger for positive returns than for negative lags. However, we do observe that LLM forecasts exhibit the same tendency to react more strongly to distant negative returns as with human forecasts, with coefficients on negative returns remaining significant at longer lags. In sum, LLM forecasts appear more symmetric than human forecasts but continue to emphasize distant negative returns in human ways.

Forcerank contests are geared toward predicting relative performance, so we also consider variants of Equation (1) where we set the dependent variable equal to the performance rank (predicted or realized) for the contest stock, and we also consider historical return ranks as the independent variables. Table 4 presents the results. In Specification (5), we continue to observe evidence of short-term reversals in realized performance when returns are ranked among a set of ten stocks. In contrast, both human and GPT-4o forecasts show strong positive extrapolation of past performance rank. While Forcerank scores load on several lags, GPT-4o performance ranks are only significant for two lags, with the most recent week's performance rank weighing heavily and helping lead to an R-squared of 0.84.

4.1.3. Extrapolation of "Cash Flows"

To assess whether LLMs extrapolate specifically from stock returns or apply similar reasoning to other types of financial data, we rerun the contest prompts but instead relabel the data as quarterly changes in scaled cash flows. We describe cash flows using language similar to Bouchaud et al. (2019), defining them as net cash flow from operating activities normalized by total assets. Importantly, the underlying numeric data remain unchanged, still based on actual weekly stock returns, so any observed differences would reflect how the LLM interprets the task based on prompt wording. The specific prompt we use is as follows:

> *The following is data on changes in cash flows for ten stocks from quarter t-12 to quarter t-1: Cash flow changes are the difference between the net cash flow from a firm's operating activities, normalized by lagged total assets. Based on the information, please rank the cash flow of these ten stocks in quarter t. How confident are you about the ranking?*

Your output will be in JSON format with the following format:
{"rank":{"1":"stock id","2":"stock id",...,"10":"stock id"}, "confidence": }
1 stands for the highest cash flow and 10 for the lowest cash flow. Confidence represents a probability that ranges from 0 to 1.

We repeat the forecasting exercise using this alternate prompt and estimate the same extrapolation regressions (Specifications (2) through (4) of Table 2). As reported in Table IA4 of the Internet Appendix, the coefficients on lagged inputs are very similar from those in the baseline specification that uses return terminology. The loadings remain strongly positive, especially for the most recent lag, and decline monotonically with lag length.

The results suggest that the extrapolative behavior observed in LLM forecasts is not unique to the labeling of the input as "returns." This indicates that LLMs apply similar reasoning across different types of financial variables and that the overweighting of recent observations is a general feature of how they process sequential numerical inputs, regardless of the specific label applied.

4.1.4. Controlling for Firm Characteristics

Financial data providers often provide many stock characteristics when investors research a company. We next consider whether including firm characteristics in the prompt influences the extent to which LLM performance forecasts extrapolate from lagged returns. We consider two sets of information: (1) Market Information, which includes Shares Outstanding, Market Capitalization, Open Price, Close Price, Day High, Day Low, 52-week high, 52-week low, 10-Day Average Volume, and Beta; and (2) Fundamental Information, including Revenue, EBITDA, Earnings Per Share, Gross Margin, Net Margin, ROE, Debt-to-Equity, P/E Ratio, and Book-to-Market ratio.

We run three additional prompts for each contest where we include, alongside the l2 weeks of lagged stock returns for each stock, either Market Information, Fundamental Information, or both. We then repeat the regressions in Equation (1). The results are tabulated in Table IA5 in the Internet Appendix. We continue to observe strong extrapolation of past returns. For example, the loading on the first lagged return is 37.45 when including Market Information for each stock in the prompt, 36.23 when including Fundamental Information, and 36.73 when including both sets of information. Across specifications, the coefficients continue to display a general monotonic decline in magnitude with each additional time lag.

4.1.5. Exponential Decay Model

The results from the linear regressions in Equation (1) indicate a clear and robust decay pattern in the relation between human and LLM performance ranks and recent past returns. To capture this pattern parsimoniously, we next estimate a parametric regression model that assumes an exponential decay of weights on past returns as follows:[18]

$$\mathrm{Y}_{\mathrm{i,t+1}} = 5.5 + \lambda_1 \cdot \sum_{s=0}^{n} w_s \mathrm{Return}_{i,t-s} + \epsilon_{i,t}, \tag{3}$$

where $w_s = \frac{\lambda_2^s}{\sum_{j=0}^{n} \lambda_2^j}$ and $\mathrm{Y}_{i,t+1}$ is either the Forcerank score or the GPT-4o forecast rank.

The first parameter (λ_1) is a scaling factor that multiplies all past returns of stock *i* and captures the level effect, i.e., the overall extent to which investor expectations respond to past returns. The second parameter (λ_2) captures the slope effect that governs how past returns are relatively weighted in forming expectations, with a λ_2 closer to zero meaning that investors put higher weight on recent past returns as opposed to distant past returns. A higher λ_1 and a lower λ_2 jointly lead to a higher degree of extrapolation, resulting in the degree of extrapolation measure $\lambda_1(1 - \lambda_2)$.

---

[18] In addition to DHJ, Greenwood and Shleifer (2014), Barberis et al. (2015), and Cassella and Gulen (2018) have also used this approach.

Table 5 presents the regression estimates for the exponential decay model. The level coefficient (λ_1) for GPT-4o forecast ranks is larger than that for humans. Additionally, the slope coefficient (λ_2) is smaller for GPT-4o, indicating a greater emphasis on recent past returns and a higher degree of extrapolation compared to human forecasts (38.03 vs 12.19 for humans). Similar results are observed for GPT-4o ranks generated based on 24 lagged weekly returns.

4.1.6. Comparing Forecasted and Realized Cross-Sectional Performance

The evidence that LLM forecasts extrapolate recent returns, combined with the evidence of short-term return reversals, suggests that GPT-4o forecasts may negatively predict future returns. On the other hand, the linear extrapolation model in Table 2 explains only 38% of the variation in GPT-4o forecasts, and it is possible that non-extrapolating aspects of LLM forecasts may be positively related to returns.

We explore the relations between forecasts and future returns using Fama-MacBeth regressions, in which the dependent variable is the daily return of an individual stock over the next week. To better understand the source of return predictability, as in DHJ, we decompose LLM and human forecasts into two components: a predicted score and the residual. The predicted score is computed as the fitted value from the regression in Equation (3); the remainder is labeled Residual GPT-4o. Predicted and residual human Forcerank scores are computed analogously. We consider a specification with the following set of firm controls that have been shown to forecast future stock returns: log market capitalization, log book-to-market, asset growth, gross profits-to-assets, market beta, weekly turnover, and the max daily return in the last month. All control variables are measured in the month of week t and prior to the return in t+1.

Table 6 reports these regression results. We find evidence that both human and LLM forecasts are negatively associated with future returns, and the evidence is most robust for

Predicted Forecasts, indicating that the extrapolative aspect of human and LLM forecasts is most negatively predictive of performance. The return predictability evidence is consistent with the view that training on human output can result in LLM return forecasts that predict future returns with the wrong sign.

4.2. LLM Extrapolation from Price Charts

Large language models can interpret images as well as numerical data, and we next examine how GPT-4o translates price charts into performance forecasts. In particular, rather than provide 12 or 24 lagged return observations in the prompt of cross-sectional performance expectations, we instead provide 10 price chart images for the contest stocks. To prevent the price level from influencing forecasts, we rescale each stock price to be 100 at the start of the price chart interval. The prompts include versions with either 12- or 24-week charts.

We construct candlestick price charts that are similar to those available from financial data providers, which in turn are likely to be included in the training data. Specifically, the charts include daily high and low values and are labeled green (red) if the day's return was positive (negative). Figure 1 depicts examples of price chart images that we provide to GPT-4o while prompting for next-week performance ranks. We then repeat the extrapolation regression using price chart-based performance rank forecasts.

The results are presented in Table 7. As in Equation (1), the regressions consider weekly returns for each stock, which may be challenging to infer from the daily candlestick price charts. Therefore, perhaps unsurprisingly, the magnitudes of the lagged return coefficients in Table 7 are smaller for the price charts than for the numerical return-based forecasts in Table 2. However, the coefficients continue to display a near monotonic downward trend in coefficient size over more distant lags. The findings are similar when using 24-week price charts as inputs for the forecasts, as reported in Table IA6. The price chart evidence suggests that the inclination for LLMs to extrapolate from recent returns extends beyond numeric data to unstructured image data.

4.3 Extrapolation in Aggregate Market Sentiment

We next turn to market-level return forecasts. An aggregate setting provides a useful complement to the stock-level analysis for two reasons. First, the AAII survey asks individual investors to forecast six-month market returns, rather than short-horizon stock returns, providing an additional performance benchmark to compare against. Second, the market context involves a single data series, which allows us to more cleanly implement robustness checks and explore alternative prompting approaches. We compare human and LLM sentiment measures by estimating the following regression model:

$$\text{Sentiment}_{i,t+1} = \gamma_0 + \sum_{s=0}^{n} \beta_s \cdot R_{i,t-s} + \varepsilon_{i,t}. \tag{4}$$

For human sentiment, $\text{Sentiment}_{i,t+1}$ for stock *i* in month *t* is a number that ranges from -1 to 1 and captures the percentage of individual investors that expect the stock market will go up over the next 6 months less the fraction of investors that think the stock market will go down.[19] For GPT-4o, $\text{Sentiment}_{i,t+1}$ is equal to 1 if the LLM assesses, based on historical returns, that the direction of the stock market over the next six months will be up, 0 if expecting no change, and -1 if expecting down. $R_{i,t-s}$ denotes S&P 500 index returns.

The regression results are presented in Table 8. Consistent with human expectations (Specification (1)), we find that GPT-4o return forecasts place larger positive weights on recent returns. In particular, the largest coefficient is on the first return lag, and the coefficients decline monotonically for the next several months. In Specification (5), we examine the extent to which

[19] Neutral is also an option, i.e., bear does not equal (1–bull).

future six-month returns are related to lagged monthly returns. None of the coefficients are significant, suggesting again that LLMs overextrapolate past monthly returns when forecasting returns over the following 6-months.

4.4 The Role of Prompt Engineering

We next explore whether prompt engineering influences the extent to which LLMs extrapolate from past returns when tasked with assessing future aggregate market performance. In particular, after the baseline query “Do you feel the direction of the stock market over the next six months will be up (bullish), no change (neutral), or down (bearish)?” we include additional instructions designed to make the sentiment forecast less extrapolative. Motivated by evidence that chain of thinking improves LLM performance (e.g., Wei, et al., 2022), in the first variant we add the instruction “Think step by step when creating your response.” We next consider a prompt variant that encourages the LLM to be more analytical, adding “Analyze the data and consider an appropriate model to apply for your response.”

We also consider two prompts that directly attempt to mitigate behavioral biases. In the first version, we address behavioral biases in general terms by adding “Consider human biases that may affect responses to similar questions and avoid these biases when creating your response.” In our final prompt, we include context-specific guidance by including a summary of Greenwood and Shleifer (2014) into the prompt as follows:[20]

> *The CSV data contain the monthly stock returns in months t−12 to t−1.*
> *Please answer the following questions:*
> *(1) Based on the return pattern, do you expect the direction of the stock market over the next six months to be up (bullish), no change (neutral), or down (bearish)?*
> *(2) How confident are you in this prediction?*

[20] The prompt used to generate the revised prompt based on the results in Greenwood and Shleifer (2014) is reported in Table IA7 in the Internet Appendix.

Before responding, consider the following empirical insight:
Investor expectations tend to be extrapolative—that is, investors often expect recent return trends to continue. However, such expectations are negatively correlated with model-based expected returns and are typically not good predictors of actual future returns. To mitigate this bias, consider whether the recent return sequence reflects a pattern likely to reverse (mean reversion) or persist.
Your output should be in JSON format with the following structure:
{"prediction": , "confidence": } where: prediction = 1 for bullish, 0 for neutral, and -1 for bearish. confidence is a probability from 0 to 1 indicating your belief in the accuracy of the prediction.

Table 9 presents the results after re-estimating Equation (4) with the additional prompts. We observe that the additional instructions have modest influence on the extent to which LLM forecasts extrapolate from past returns. Encouraging the LLM to apply a model produces a loading on the first lagged monthly return that is somewhat smaller, 7.61, compared to the baseline of 9.17. The largest decline in extrapolation occurs when including the summary of Greenwood and Shleifer (2014), where the first-lag coefficient falls to 6.21 in Specification (4). However, the loadings continue to be positive in all of the specifications, and they decline roughly monotonically for each of the prompts.

The findings provide some support for the view that LLM biases can be reduced through carefully crafted prompts. On the other hand, overextrapolation remains challenging to fully eliminate. Moreover, generic instructions aimed at bias removal tend to have little impact, which suggests that understanding specific LLM biases in advance is necessary for effective mitigation.

4.5 Alternative Large Language Models

Our analysis thus far has relied on forecasts generated by the GPT-4o large language model from OpenAI. In this section, we consider market sentiment measures generated using alternative large language models.

We repeat the market sentiment queries using other LLMs that are widely regarded at the time of the analysis: the Claude 3.5 Sonnet model from Anthropic, and the Gemini 2.5 Pro model from Google. We also consider the most recent model by OpenAI, released in May of 2024, which is labeled o1 preview. Table IA8 in the Internet Appendix presents the results. We note that the different LLM sentiment forecasts are closely related. In particular, the correlations between GPT-4o sentiment and the sentiment measures generated by GPT-o1 preview, Claude 3.5 Sonnet, and Gemini 2.5 Pro are 0.71, 0.74, and 0.67, respectively. Moreover, the magnitudes of the extrapolative coefficients are very similar when comparing Specification (2) of Table 8 to extrapolative coefficients in Table IA8.

4.6 Sensitivity of LLM Extrapolation to Underlying Autocorrelation

Afrouzi et al. (2023) finds experimental evidence that humans overreact to the most recent observation when forming expectations of a stable random series, and that overreaction is stronger for less persistent processes. Bordalo et al. (2020) find supporting evidence in a sample of professional forecasters, with evidence of overreaction to macro news that is stronger for less persistent series. We conjecture that LLM forecasts may be similarly influenced by the level of autocorrelation present in the data provided in the prompt.

Specification (5) of Table 8 indicates little evidence of autocorrelation in aggregate market returns, yet it is possible that LLM extrapolation may vary for series with high or low persistence. We explore this hypothesis by constructing simulated autoregressive market returns that match the mean and standard deviation of actual returns. In particular, we simulate 12 monthly returns using an AR(1) process that matches the mean and standard deviation of the market's monthly returns over the preceding 12 months. We do this for each of the 1,164 months from January 1927 to

December 2023 and for seven different autocorrelation values: -0.9, -0.6, -0.3, 0, 0.3, 0.6, and 0.9. We then repeat the GPT-4o market sentiment prompts using the simulated return data.

Table IA9 in the Internet Appendix reports the regression results based on Equation (4) using the simulated return data with different levels of autocorrelation. We observe that the degree of extrapolation by LLMs remains remarkably consistent regardless of the level of autocorrelation. For example, the coefficient on the first lagged return is 6.6 when the correlation is 0.9, 6.4 when the autocorrelation is 0.0, and 5.9 when the autocorrelation is -0.9. The coefficients on all of the lags are significantly positive regardless of the autocorrelation structure provided in the prompt. Comparing the distance between the predicted persistence to the actual level of persistence suggests that LLMs overreact to recent information more when the series is less persistent, consistent with human behavior.

Together, the evidence indicates that return overextrapolation in LLM forecasts is not (i) confined to individual stocks or a specific large language model, (ii) easy to eliminate through simple prompt engineering, or (iii) sensitive to the level of underlying autocorrelation in the data.

4.7 How Do LLMs Justify Their Forecasts?

To better understand the reasoning behind LLM extrapolation behavior, we append the baseline aggregate forecast prompt with a follow-up instruction: "*Please explain how you arrived at this prediction*." This addition was applied across the 438 monthly forecast prompts in which GPT-4o was provided with historical market return data and asked to generate a six-month-ahead sentiment forecast. The resulting responses contained a wide range of textual justifications, often referencing specific return patterns or broader market sentiment. To systematically summarize the common themes in these explanations, we issued a second prompt:

> *Your primary task is to analyze provided textual justifications for stock market predictions, which may span different dates or contexts. Carefully read through all the input reasons to identify the core drivers and recurring themes. From these justifications, extract and summarize the most significant and distinct reason that underpin the stock market predictions. Ensure your summary is based strictly on the information provided in the input, without introducing external knowledge or speculation. Please summarize succinctly.*

The resulting rationale summary is as follows: "The primary reason for the stock market predictions across the provided analyses is the identification of recent positive trends in stock returns. Many analyses highlight a pattern of positive returns in recent months, suggesting upward momentum and a bullish sentiment. This is often supported by the magnitude of positive returns outweighing negative ones, indicating strong market performance and investor confidence. Despite some volatility and mixed signals, the consistent positive returns in recent months are seen as a sign of potential continued growth, leading to a generally bullish outlook with moderate confidence."

This simple textual analysis complements the regression evidence, showing that the tendency to extrapolate from recent positive returns also appears in the LLM's own stated reasoning. The explanations are similar to common trend-following heuristics and suggest that the model draws on the types of reasoning frequently found in financial commentary. While the responses do not reflect deeper reasoning about mean reversion or risk, they offer a consistent interpretation of the data, indicating that overextrapolation is both systematic and explicitly justified by the model.

5. Bias and Miscalibration in Large Language Model Forecasts

Ben-David, Graham, and Harvey (2013) document that CFOs tend to be overconfident in their market return forecasts, with realized outcomes frequently falling outside the stated 80%

confidence intervals. More recent work by Hartzmark and Sussman (2024) confirms that human forecasts often fail to capture the full range of possible outcomes, highlighting persistent miscalibration in expectations about market returns. Motivated by these findings, we assess whether similar patterns arise when GPT-4o is prompted to generate distributional forecasts based on historical return data.

5.1 LLM Expected Return Forecast Accuracy

We begin by examining whether LLM expected return forecasts appear biased relative to realized outcomes. Humans tend to be overly optimistic in a variety of settings. For example, Van den Steen (2004) argues that investors are likely to choose stocks for which they have overestimated the likelihood of success, similar to the winner's curse, and Hartzmark, Hirshman, and Imas (2021) show that simply owning an asset makes investors prone to optimistic beliefs about its performance. If overoptimism manifests in the training data, then LLM forecasts may also be higher than realized returns. Table 10 Panel A presents forecast statistics for the sample of 485 stock-months that survived the historical return requirements.[21] We observe that the cross-sectional mean of the GPT-4o forecast for next month's stock return is 2.0%. The LLM forecast is considerably larger than the average historical mean for the data that were provided in the prompt (1.38%) and the magnitude of next month's realized return (1.12%), with the former being statistically significant.[22]

Figure 2 presents a scattered bin plot of the LLM return forecasts against the historical mean of the 120 monthly returns provided in each prompt, which allows us to visually assess the

---

[21] 15 prompts returned explanatory text or unusable output, so the final forecast sample includes 485 valid forecasts.
[22] When comparing the forecast to next month's realized return, the p-value is 0.11 in the sample of 500 prompts, likely due to the noise in realized returns. In earlier runs using 10,000 prompts, the difference was highly significant. We reduced the number of prompts to 500 to make it affordable to run a broader variety of robustness checks.

extent to which the degree of bias varies across the distribution of input values. The upward-sloping line confirms that the LLM conditions its forecasts on the input data: higher historical means lead to higher forecasts. However, the LLM consistently overestimates expected returns relative to the historical means, as evidenced by the forecast line lying above the 45-degree line throughout. Notably, the slope of the LLM forecast line is less than one, indicating that the extent of overestimation is larger when historical means are low and diminishes as historical means rise. In other words, the LLM exhibits stronger optimism when the underlying data suggest more modest returns and becomes relatively better calibrated as the historical average increases.

Figure 3 (top panel) plots the cross-sectional distribution of the historical means of the data provided in the prompts and the distribution of resulting GPT-4o forecasts. The distribution of historical means appears roughly Gaussian, suggesting that the historical sample sizes are sufficient for the central limit theorem to apply. On the other hand, GPT-4o forecasts are decidedly less smooth. Very few of the GPT-4o expected return forecasts are below zero (0.62% of forecasts), which suggests that the LLM's training may have embedded the idea that expected returns should be nonnegative. However, Table 10 indicates that the median, 75th, and 95th percentiles of the distribution of GPT-4o expected return forecasts are all higher than the average historical equivalents, suggesting that GPT-4o's positive expected return bias extends beyond truncating at zero.[23]

5.2 LLM Return Forecast Calibration Evidence

[23] Sias, Starks, and Turtle (2024) argue that investor decision-making in the context of retirement planning is more sensitive to long-term return expectations than to short-term fluctuations. Motivated by this perspective, we prompt the LLM to forecast the "*average monthly return over the next 10 years*," using the same 120 months of historical returns as in our baseline setting. The resulting forecasted mean monthly return is 4.11%, which suggests that optimism in LLM forecasts is not restricted to short horizons.

We next examine LLM forecasts of Low and High returns. In Panel A of Table 10, the average GPT-4o 80% confidence interval is 21.1%, which is smaller than the average historical confidence interval of 24.8%. We observe (in Panel C) that next month's realized return value lies within the GPT-4o confidence interval 69.3% of the time, which is less accurate than the 78.0% that could be contained by simply using the 10th and 90th historical percentiles as the forecasts. The miscalibration is primarily on the upside, with 19.3% of realized returns occurring above the High forecast, compared to 11.4% on the downside. This pattern is evident in the bottom two panels of Figure 3. The distribution of low GPT-4o forecasts is not markedly different than the distribution of historical 10th percentiles. On the other hand, GPT-4o high forecasts appear routinely lower than historical 90th percentiles.

The evidence of miscalibration in GPT-4o forecasts is considerably less severe than in the executive surveys. BGH report that realized market returns fall within CFOs' 80% confidence intervals only 36% of the time. Hartzmark and Sussman (2024) find that participants perform better when provided with historical data but continue to exhibit bias. However, their confidence intervals demonstrate limited sensitivity to the requested level of confidence, indicating that even well-informed forecasts may underreact to changes in the implied range of outcomes.

In Figure 4, we present LLM projected forecasted confidence interval widths for ranges of 60%, 80% (baseline), 90%, and 98%, and compare them to the actual return distribution implied by historical data. Relative to human responses documented in Hartzmark and Sussman (2024), the LLM appears generally better calibrated, with forecasted intervals expanding appropriately as the confidence level increases and more closely matching the actual structure of return distributions.

On the other hand, GPT-4o's confidence intervals are consistently narrower than those implied by the historical data, especially for wider confidence intervals. This discrepancy is primarily driven by a downward bias in the upper bound of the forecasted range, as evident in the bottom panel of Figure 3 and Figure 4. The evidence suggests that while LLM forecasts incorporate the idea that more extreme outcomes are associated with wider intervals, they underestimate the positive skewness of individual stock returns. This downward compression is less pronounced at lower confidence levels, such as the 60% interval, where the effect of skewness is naturally diminished.

We next consider whether prompt engineering can reduce the bias observed in LLM expected return forecasts. Building on the extrapolation analyses in Table 8, we experiment with four prompt variations aimed at encouraging more calibrated distributional forecasts. These include appending the prompt with the admonition to: 1) "Think step by step when creating your response;" 2) "Analyze the data and consider an appropriate model to apply for your response;" and 3) "Consider human biases that may affect responses to similar questions and avoid these biases when creating your response." In our final prompt, we include context-specific guidance by including a summary of Hartzmark and Sussman (2024) into the prompt:[24] The revised prompt is designed to nudge the LLM toward constructing a more deliberate belief distribution. It reads:

> *You are presented with monthly returns for a financial asset over the past 120 months. Using this information, estimate the belief distribution for next month's return by first considering a full probability distribution of possible outcomes. Based on this distribution, provide:*
> *The 10th percentile return (i.e., a 1-in-10 chance the return is lower than this value): a%. The expected return (mean of the distribution): b%. The 90th percentile return (i.e., a 1-in-10 chance the return is higher than this value): c%*
> *Return your answer as a JSON object: {"low": a%, "expected": b%, "high": c%}*

[24] The prompt used to generate the revised prompt based on the results in Greenwood and Shleifer (2014) is reported in Table IA10 in the Internet Appendix.

Before answering, construct a mental histogram of possible outcomes across bins (e.g., from –20% to +20%) to avoid biases in point estimates and confidence intervals. Be mindful that directly reporting means and intervals can introduce upward bias and underestimate or misstate uncertainty. Reflect on the entire return distribution before summarizing.

Table 11 presents the results. Across all four prompt variants, expected return forecasts remain biased upward by 56 to 87 basis points relative to the historical mean, similar to the 61. Calibration remains relatively poor as well, with the upper bound (high forecast) continuing to fall well below the 90th percentile of the historical data across all specifications. Despite explicit instructions to reason carefully, apply models, or avoid behavioral bias, the LLM's distribution forecasts continue to display optimistic bias and understate the right tail of the return distribution. The evidence suggests that straightforward prompt engineering is insufficient to eliminate the upward bias and systematic miscalibration in LLM expected return forecasts.

Hartzmark and Sussman (2024) find evidence that prompting individuals to provide probability forecasts across return bins, rather than point forecasts, reduces the optimistic bias in expectations. In particular, they infer expected returns from the bin distributions and find these are better aligned with the true mean of the historical data shown to participants. Motivated by this result, we adapt their approach to the LLM setting. Specifically, we prompt GPT-4o with historical monthly return data for a randomly selected stock-month and ask it to assign probabilities that the next month's return will fall within a set of predefined return bins ranging from –30% to +30%. The specific prompt is reported in Table IA11.

The resulting bin probabilities are summarized in Figure 5, which compares the average forecasted distribution from GPT-4o to the empirical return distribution from the historical data provided in each prompt. In contrast to the findings in Hartzmark and Sussman (2024), we do not find that prompting the LLM for bin-level forecasts reduces bias. The LLM forecasts assign higher likelihoods to positive return bins than are observed in the historical data, suggesting continued optimism. Thus, while distributional elicitation may help debias human forecasts, it does not appear to mitigate optimistic bias in LLM-generated return distributions.

5.3 Framing Effects in Price and Return Forecasts

Glaser, Iliewa, and Weber (2019) (GIW) show that participants presented with identical historical information respond with systematically different return expectations depending on how the information is presented and the task is framed. In particular, asking for a return percentage rather than a price level leads to more optimistic expectations, while displaying accompanying data as a return bar chart rather than a price line chart results in more pessimistic return forecasts. The findings suggest that both the format of the question and the visual presentation of information can meaningfully influence human expectations.

In this section, we prompt LLMs to provide both price and return forecasts under comparable conditions. Using the same 500 stock-months from the previous section, we generate price-level line charts and return bar charts covering the most recent 12 months, closely following the visual format used in GIW (example charts are shown in Figure 6). For each chart type, i.e., price line or return bar, we prompt GPT-4o to forecast either the asset's price level at the end of the next month or the return over the next month. Below, we present two of the four prompt variations to illustrate how the wording differs across tasks.

> *Below is the price-level chart for a financial asset covering the last 12 months, including the current price in monetary units (MU): What percentage return do you expect the asset to deliver over the next month? Return only a JSON object in this exact schema: {"expected_return": <numeric %>}*

> *Below is a return-percentage chart for a financial asset covering the last 12 months, including the current price in monetary units (MU): What price level (MU)*

do you expect the asset to close exactly one month from now? Return only a JSON object in this exact schema: {"expected_price": <numeric>}

Figure 7 displays the average expected returns from LLM forecasts across the four different combinations of visual stimulus (price chart vs. return chart) and task framing (forecast price vs. forecast return), along with 95% confidence bounds. While GIW find that human expectations are influenced by both the task (with return forecasts more optimistic than returns inferred from price forecasts) and the stimulus (with return charts leading to less optimistic expectations than price charts). We find that LLM forecasts exhibit a different pattern.

While the framing of the task (predicting price vs. return forecast) has little discernible effect on average expectations, the way past data are visually displayed plays a stronger role. Specifically, when prompted with price charts, GPT-4o generates return forecasts that fall significantly below the historical mean included in the prompt. On the other hand, when shown return charts, GPT-4o forecasts are above the mean, consistent with the earlier findings using return data directly (i.e., not visually presented in a bar chart). The framing effect is significant. In a regression of GPT-4o return forecast on task and display indicator variables, the coefficient on the return-chart indicator is 1.76% ($t = 10.47$). These results suggest that LLM return forecasts are influenced more by how historical information is visually represented than by the nominal framing of the forecast task.

5.4 Discussion

Our analysis suggests that LLM stock performance forecasts exhibit overextrapolative behavior, tend to be overoptimistic regarding expected returns, and are downward-biased when forecasting the right tail of the return distribution. A natural question that arises is whether it is easy to "turn off" these biases. We contend that completely removing behavioral biases from

LLMs will be difficult. The issue is not that LLMs are unaware of investor biases. For example, when asked, "What behavioral biases do investors make when using historical returns to predict future returns," GPT-4o's response includes Extrapolation Bias, Recency Bias, Overconfidence, Confirmation Bias, Hindsight Bias, and the Availability Heuristic, and it offers definitions of each. GPT-4o can also easily summarize the evidence on short-term return reversals.

The challenge is that these biases are deeply rooted in the training process. LLMs are trained on vast datasets that reflect the full spectrum of human thought, including the biases and heuristics that are prevalent in financial discussions. The volume of data required to train an LLM makes it difficult to eliminate all instances of bias without impairing the model's ability to generate coherent and contextually appropriate responses. Consequently, even if an LLM "understands" what these biases are in theory, its output may still reflect these biases because they are embedded in the data from which the model learns. We find that additional instructions to "use a model" or "avoid biases" have little effect on LLM output.

LLMs that are pre-trained on a broad corpus can be fine-tuned using more targeted datasets, improving performance on domain-specific tasks. Repeated interactions can also reduce bias, as the model adapts to user-specific prompting styles that emphasize rational, statistical reasoning. However, in the absence of such usage history, and particularly when an LLM faces a task without well-answered examples in the fine-tuning data, LLMs are likely to fall back on general knowledge from training data, which may embed common biases and heuristics.

On one hand, designing Generative AI models to mimic human thought and emotion can have beneficial applications in personal finance contexts. For example, human-like LLMs can help researchers explore which financial nudges are likely to be welfare-enhancing (e.g., Beshears et al., 2015; 2017). Researchers have begun using LLM personas to approximate the behavior of human macroeconomic forecasters (Hansen, et al., 2025) and to capture differential reactions to news (Bhagwat, et al., 2025).

In contrast, when generative AI models are used in expert roles, such as financial advising, it is crucial that they maintain neutrality, rationality, and provide data-driven insights. Expert contexts demand a systematic removal of bias to ensure informed and objective outputs. Currently, LLMs blend human-like and expert characteristics, reflecting the complexities of their training. Our findings highlight the importance of understanding how these models capture aspects of human behavior and emphasize the need for targeted development to shape LLMs for specific roles.

6. Conclusion

AI's capacity for objectively analyzing vast amounts of information has the potential to revolutionize financial decision-making. However, large language models (LLMs) and other AI algorithms are trained on human output, which raises the risk that they may inherit cognitive biases embedded in the data. This study examines whether OpenAI's GPT-4o exhibits patterns consistent with interpreting historical return data and price chart images to forecast stock returns. In particular, we assess how LLMs respond to the timing, magnitude, and presentation of past returns when forming expectations about future outcomes.

Our empirical analysis indicates that GPT-4o and human forecasts rely on past data in similar ways, with a positive, gradually declining (over)emphasis on lagged returns. This pattern is not present in the cross-section of realized returns, which instead tend to exhibit a pattern of short-term reversals. The behavior of LLM forecasts is consistent with documented overextrapolative expectations in human decision-making. The evidence remains robust when using alternative LLMs and also when providing price charts instead of numerical return data. To address concerns about look-ahead bias, we also conduct additional analysis using post-training samples, which confirms the persistence of these patterns. Moreover, our analysis of simulated data with varying levels of autocorrelation reveals that LLM extrapolation is consistent across different levels of persistence, highlighting a distinct and systematic bias that emerges independently of the underlying data structure.

Beyond extrapolation, we find consistent evidence of overoptimism in LLM expected return forecasts. GPT-4o predicts higher average returns than those realized or implied by historical means. While the model generally performs better than human forecasters in characterizing risk, its distribution forecasts deviate from historical benchmarks. Confidence intervals are too narrow, particularly on the upside, reflecting an underestimation of extreme positive outcomes. This compression of the right tail is even more pronounced at higher confidence levels, indicating that the LLM struggles to capture the positive skewness of stock returns. Distributional forecasts elicited through prompts regarding return bin likelihoods confirm that the upward bias persists even when asking the model to allocate probabilities across return ranges.

In our final analysis, we examine whether how the forecast task is framed and how historical information is presented meaningfully affects LLM outputs. Holding the underlying data constant, we find that task framing (forecasting prices vs. returns) has little effect on return forecasts, whereas the visual stimulus matters. Prompting GPT-4o with price charts elicits significantly lower return expectations than return bar charts, suggesting that model training processes price extrapolation differently than return extrapolation.

We note that our results should be interpreted with caution, given several important limitations. LLM outputs are sensitive to prompt framing and may vary across users depending on prior interactions and interface differences. In addition, the models we study are trained on a broad and evolving corpus, which introduces variability that can be difficult to control. While future research may explore whether fine-tuning LLMs on financial forecast data can produce outputs better suited to specific forecasting goals, we emphasize the importance of understanding how off-the-shelf models behave, particularly in light of their growing influence on investor decision-making.

More broadly, our findings contribute to the literature on AI integration in finance. Although LLMs show promise in characterizing risk and producing useful forecasts, they also exhibit a blend of statistical reasoning and human-like behavioral biases. These patterns are context-dependent and often persist despite efforts to guide or restructure model responses. As LLMs become more widely adopted in financial tools, recognizing how and when these biases arise will be essential for their effective use.

References

Afrouzi, H., Kwon, S.Y., Landier, A., Ma, Y. and Thesmar, D., 2023. Overreaction in Expectations: Evidence and Theory. *The Quarterly Journal of Economics*, 138: 1713-1764.

Argyle, L.P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C. and Wingate, D., 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31: 337-351.

Atmaz, A., Gulen, H., Cassella, S. and Ruan, F., 2024. Contrarians, Extrapolators, and Stock Market Momentum and Reversal. *Management Science*, 70: 5949-5984.

Avramov, D., Chordia, T. and Goyal, A., 2006. Liquidity and Autocorrelations in Individual Stock Returns. *The Journal of Finance*, 61(5), 2365-2394.

Barberis, N., 2018. Psychology-Based Models of Asset Prices and Trading Volume. In *Handbook of Behavioral Economics: Applications and Foundations* 1 (Vol. 1: 79-175). North-Holland.

Barberis, N., R. Greenwood, L. Jin, and A. Shleifer. 2015. X-CAPM: An Extrapolative Capital Asset Pricing Model. *Journal of Financial Economics* 115: 1-24.

Bartlett, R., Morse, A., Stanton, R. and Wallace, N., 2022. Consumer-Lending Discrimination in the FinTech Era. *Journal of Financial Economics*, 143: 30-56.

Ben-David, I., Graham, J.R. and Harvey, C.R., 2013. Managerial Miscalibration. *The Quarterly Journal of Economics*, 128: 1547-1584.

Bereska, L. and Gavves, E., 2024. Mechanistic Interpretability for AI Safety–A Review. arXiv preprint arXiv:2404.14082.

Bertomeu, J., Lin, Y., Liu, Y. and Ni, Z., 2025. The Impact of Generative AI on Information Processing: Evidence from the Ban of ChatGPT in Italy. *Journal of Accounting and Economics*, p.101782.

Beshears, J., Choi, J.J., Laibson, D., Madrian, B.C. and Milkman, K.L., 2015. The Effect of Providing Peer Information on Retirement Savings Decisions. *The Journal of Finance*, 70: 1161-1201.

Beshears, J., Choi, J.J., Laibson, D. and Madrian, B.C., 2017. Does Aggregated Returns Disclosure Increase Portfolio Risk Taking? *The Review of Financial Studies*, 30: 1971-2005.

Bhagwat, V., Cookson, J.A., Dim, C. and Niessner, M., 2025. The Market's Mirror: Revealing Investor Disagreement with LLMs. Available at SSRN 5375473.

Bini, P., Cong, L.W., Huang, X. and Jin, L.J., 2025. Behavioral Economics of AI: LLM Biases and Corrections. Available at SSRN 5213130.

Bordalo, P., Gennaioli, N., Ma, Y. and Shleifer, A., 2020. Overreaction in Macroeconomic Expectations. *American Economic Review*, 110: 2748-2782.

Bouchaud, J.P., Krueger, P., Landier, A. and Thesmar, D., 2019. Sticky Expectations and the Profitability Anomaly. *The Journal of Finance*, 74: 639-674.

Bowen III, D.E., Price, S.M., Stein, L.C. and Yang, K., 2024. Measuring and Mitigating Racial Bias in Large Language Model Mortgage Underwriting. Available at SSRN 4812158.

Bybee, J., 2023. The Ghost in the Machine: Generating Beliefs with Large Language Models. arXiv preprint arXiv:2305.02823.

Cao, S., Jiang, W., Wang, J. and Yang, B., 2024. From Man vs. Machine to Man + Machine: The Art and AI of Stock Analyses. *Journal of Financial Economics*, 160, p.103910.

Cassella, S., and H. Gulen. 2018. Extrapolation Bias and the Predictability of Stock Returns by Price-Scaled Variables. *The Review of Financial Studies* 31(11), 4345–4397.

Cassella, S., Chen, Z., Gulen, H. and Petkova, R., 2023. Extrapolators at the Gate: Market-Wide Misvaluation and the Value Premium. *Available at SSRN 3705481*.

Chen, H., Didisheim, A. and Somoza, L., 2025. Out of the Black Box: Uncertainty Quantification for LLMs via Conditional Probabilities. Available at SSRN 5012852.

Chen, Y., Liu, T.X., Shan, Y. and Zhong, S., 2023. The Emergence of Economic Rationality of GPT. *Proceedings of the National Academy of Sciences*, 120: 2316205120.

Chen, S., Peng, L. and Zhou, D., 2024. Wisdom or Whims? Decoding Investor Trading Strategies with Large Language Models. *Available at SSRN 4867401.*

Chen, J., Tang, G., Zhou, G., and Zhu, W., 2025. ChatGPT and DeepSeek: Can They Predict the Stock Market and Macroeconomy? Available at SSRN 4660148.

Choi, J.J., Huang, D., Yang, Z. and Zhang, Q., 2024. Better than Human? Experiments with AI Debt Collectors. Working Paper, Yale University.

Chui, A.C., Subrahmanyam, A. and Titman, S., 2022. Momentum, Reversals, and Investor Clientele. *Review of Finance*, 26: 217-255.

Da, Z., Huang, X. and Jin, L., 2021. Extrapolative Beliefs in the Cross-section: What Can We Learn from the Crowds? *Journal of Financial Economics*, 140: 175-196.

Da, Z., Liu, Q. and Schaumburg, E., 2014. A Closer Look at the Short-Term Return Reversal. *Management Science*, 60: 658-674.

Erel, I. Stern, L., Tan, C., and Weisbach. M., 2021. "Selecting Directors Using Machine Learning." *The Review of Financial Studies,* 34: 3226–3264.

Fedyk, A., Hodson, J., Khimich, N. and Fedyk, T., 2022. Is Artificial Intelligence Improving the Audit Process? *Review of Accounting Studies*, 27: 938-985.

Fedyk, A., Kakhbod, A., Li, P. and Malmendier, U., 2024. AI and Perception Biases in Investments: An Experimental Study. Available at SSRN 4787249.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T. and Walther, A., 2022. Predictably Unequal? The Effects of Machine Learning on Credit Markets. *The Journal of Finance*, 77: 5-47.

Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., and Ahmed, N.K., 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, pp.1-79.

Glaser, M., Iliewa, Z. and Weber, M., 2019. Thinking about Prices versus Thinking about Returns in Financial Markets. *Journal of Finance*, 74: 1249-1283.

Glasserman, P. and Lin, C., 2023. Assessing Look-Ahead Bias in Stock Return Predictions Generated by GPT Sentiment Analysis. arXiv preprint arXiv:2309.17322.

Greenwood, R. and Shleifer, A., 2014. Expectations of Returns and Expected Returns. *The Review of Financial Studies*, 27: 714-746.

Gulen, H. and Lim, C., 2024. Decoding Expectation Formation from Realized Stock Prices: An Eye-Tracking Study. Available at SSRN 4610951.

Hansen, A.L., Horton, J.J., Kazinnik, S., Puzzello, D., and Zarifhonarvar, A., 2025. Simulating the Survey of Professional Forecasters. Available at SSRN 5066286.

Hartzmark, S.M., Hirshman, S.D. and Imas, A., 2021. Ownership, Learning, and Beliefs. The *Quarterly Journal of Economics*, 136: 1665-1717.

Hartzmark, S.M. and Sussman, A.B., 2024. Eliciting Expectations. Available at SSRN 4780506.

Henning, T., Ojha, S.M., Spoon, R., Han, J. and Camerer, C.F., 2025. LLM Trading: Analysis of LLM Agent Behavior in Experimental Asset Markets. arXiv preprint arXiv:2502.15800.

Hirshleifer, D., 2015. Behavioral Finance. In *Annual Review of Financial Economics*, 7: 133-159.

Horton, J.J., 2023. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? (No. w31122). National Bureau of Economic Research.

Jegadeesh, N., 1990. Evidence of Predictable Behavior of Security Returns. *The Journal of Finance*, 45: 881-898.

Jha, M., Qian, J., Weber, M. and Yang, B., 2024. ChatGPT and Corporate Policies (No. w32161). National Bureau of Economic Research.

Jiang, J., Kelly, B. and Xiu, D., 2023. (Re-) Imag (in) ing Price Trends. *The Journal of Finance*, 78: 3193-3249.

Kim, A., Muhn, M. and Nikolaev, V.V., 2024. Bloated Disclosures: Can ChatGPT Help Investors Process Information? Chicago Booth Research Paper.

Kuchler, T. and Zafar, B., 2019. Personal Experiences and Expectations about Aggregate Outcomes. *The Journal of Finance*, 74: 2491-2542.

Kuhnen, C.M., 2015. Asymmetric Learning from Financial Information. *The Journal of Finance*, 70: 2029-2062.

Kuhnen, C.M. and Knutson, B., 2005. The Neural Basis of Financial Risk Taking. *Neuron*, *47*: 763-770.

Lehmann, B. N, 1990. Fads, Martingales, and Market Efficiency. *The Journal of Financial Economics*, 105: 1-28.

Li, P., Castelo, N., Katona, Z. and Sarvary, M., 2024. Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis. *Marketing Science*, 43: 254-266.

Lim, Y., 2024. Is Artificial Intelligence (AI) Risk-Averse? Working Paper, Texas Woman's University.

Lohr, Steve, 2024. A.I. Can Write Poetry, but It Struggles With Math. *The New York Times*, July 23. Available at https://www.nytimes.com/2024/07/23/technology/ai-chatbots-chatgpt-math.html

Lopez-Lira, A. and Tang, Y., 2023. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. arXiv preprint arXiv:2304.07619.

Lopez-Lira, A., Tang, Y., and Zhu, M., 2025. The Memorization Problem: Can We Trust LLMs' Economic Forecasts? Available at SSRN 5217505.

Ludwig, J., Mullainathan, S., and Rambachan, A., 2025. Large Language Models: An Applied Econometric Framework. National Bureau of Economic Research, No. w33344.

Reher, M. and Sokolinski, S., 2024. Robo Advisors and Access to Wealth Management. *Journal of Financial Economics*, 155: 103829.

Ross, J., Kim, Y. and Lo, A.W., 2024. LLM Economicus? Mapping the Behavioral Biases of LLMs via Utility Theory. arXiv preprint arXiv:2408.02784.

Sarkar, S.K. and Vafa, K., 2024. Lookahead Bias in Pretrained Language Models. Available at SSRN 4754678.

Satpute, A., Gießing, N., Greiner-Petter, A., Schubotz, M., Teschke, O., Aizawa, A. and Gipp, B., 2024. Can LLMs Master Math? Investigating Large Language Models on Math Stack Exchange. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2316-2320).

Sias, R., Starks, L. and Turtle, H., 2024. Long-Horizon Beliefs and Financial Choices. Working Paper, University of Texas.

Van den Steen, E., 2004. Rational Overoptimism (and Other Biases). *American Economic Review*, 94: 1141-1151.

Van, P. H., and Cunningham, S., 2024. Can Base ChatGPT Be Used for Forecasting Without Additional Optimization? Available at SSRN 4907279.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V. and Zhou, D., 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35: 24824-24837.

Appendix A

A.1 Forecast Variables

- $\text{Forcerank}_{\text{i,t}}$ – The end-of-week-*t* consensus ranking based on investors' average expectation regarding the performance of stock *i* over week *t* + 1. The rank ranges from 1 to 10 based on the ten stocks in each Forcerank contest. Source: Estimize.
- $\text{GPT}-4\text{o Rank}_{i,t}$– GPT-4o's ranking of the performance of stock *i* for week *t* when provided with historical return data. Source: GPT-4o prompts.
  - $\text{GPT}-4\text{o Rank}_{\text{i,t}}^{12w}$ – 12 weeks of historical return data are included in the prompt.
  - $\text{GPT}-4\text{o Rank}_{\text{i,t}}^{24w}$ – 24 weeks of historical return data are included in the prompt.
  - $\text{GPT}-4\text{o Rank}_{\text{i,t}}^{12w\ Chart}$ – An image of a price chart with 12 weeks of historical return data is included in the prompt.
  - $\text{GPT}-4\text{o Rank}_{\text{i,t}}^{24w\ Chart}$ – An image of a price chart with 24 weeks of historical return data is included in the prompt.
- $\text{Predicted GPT}-4\text{o Rank}_{i,t}$ – The fitted value obtained from regressing GPT-4o rank on lagged returns as in Specification (2) of Table 5
- $\text{Residual GPT}-4\text{o Rank}_{i,t}$ – The residual value obtained from regressing GPT-4o rank on lagged returns as in Specification (2) of Table 5.
- $AAII\ Sentiment_t$ – The American Association of Individual Investors Bull – Bear Spread, defined as the fraction of survey respondents at the end of the last week of month *t* that feel the direction of the stock market over the next six months will be up (bullish), less the fraction of survey respondents that feel the direction of the stock market over the next six months will be down (bearish). Note "no change" (neutral) is also a survey option. Source: Bloomberg.
- $ChatGPT\ Sentiment_t$ – GPT-4o's market sentiment score for month *t* when provided with 12 lagged monthly returns for the S&P 500 index and asked, "Do you feel the direction of the stock market over the next six months will be up (bullish), no change (neutral) or down (bearish)?" A score of 1 represents bullish sentiment, 0 represents neutral sentiment, and -1 represents bearish sentiment. Source: GPT-4o prompts.
- $\text{GPT}-4\text{o Low}_{i,t}$ – GPT-4o's response to the prompt "There is a 1-in-10 chance the actual return will be less than x%," for stock *i* in month *t*+1, when provided with up to ten (but no fewer than five) years of monthly stock returns. Source: GPT-4o prompts.
- $\text{GPT}-4\text{o Expected}_{i,t}$ – GPT-4o's response to the prompt "I expected the next month's return to be x%" for stock *i* in month *t*+1, when provided with up to ten (but no fewer than five) years. Source: GPT-4o prompts.
- $\text{GPT}-4\text{o High}_{i,t}$ – GPT-4o's response to the prompt "There is a 1-in-10 chance the actual return will be greater than x%," for stock *i* in month *t*+1, when provided with up to ten (but no fewer than five) years of monthly stock returns. Source: GPT-4o prompts.

A.2 Return Measures

- $\text{Return}_{i,t}$ – Return for stock *i* in week or month *t*. Source: CRSP
- $\text{Return Rank}_{i,t}$ – Stock return performance rank for stock *i* in week *t*. The rank is from 1 to 10 for the ten stocks in the Forcerank contest.
- $\text{S\&P Return}_t$ – Return for the S&P 500 Index for month *t*. Source: CRSP.

A.3 Control Variables

- *Market Capitalization*$_{i,t}$ – The market value of equity measured for month *t*. Source: CRSP.
- *Book to Market*$_{i,t}$ – The ratio of the book value of equity to the market value of equity, measured for the fiscal year prior to month *t*. Source: Compustat.
- *Asset Growth*$_{i,t}$ – The percentage change in book value of total assets from balance sheet, measured for the fiscal year prior to month *t*. Source: Compustat.
- *Profitability*$_{i,t}$ – Revenue minus cost of goods sold, divided by total assets. Measured for the fiscal year prior to month *t*. Source: Compustat.
- *Market Beta*$_{i,t}$ – Market beta from fitting the CAPM to daily stock returns for stock *i* in month *t*. Source: CRSP.
- *Return MAX*$_{i,t}$ – The maximum daily return for stock *i* in month *t*. Source: CRSP.
- *Turnover*$_{i,t}$ – The sum of daily dollar volume over market cap for stock *i* in week *t*. Source: CRSP.

A.4 Additional Variables for GPT-4o Forecasts

- *Shares Outstanding*$_{i,t}$ – Number of shares outstanding (in million) for stock i on Friday of week *t*. Source: CRSP.
- *Market Capitalization*$_{i,t}$ – The market value of equity on Friday of week *t*. Source: CRSP.
- *Open Price*$_{i,t}$ – The open price on Friday in week *t*. Source: CRSP.
- *Close Price*$_{i,t}$ – The close price on Friday in week *t*. Source: CRSP.
- *Day High*$_{i,t}$ – The highest trading price on Friday in week *t*. Source: CRSP.
- *Day Low*$_{i,t}$ – The lowest trading price on Friday in week *t*. Source: CRSP.
- *52 Week High*$_{i,t}$ – The highest daily close price in the past 52 week. Source: CRSP.
- *52 Week Low*$_{i,t}$ – The lowest daily close price in the past 52 week. Source: CRSP.
- *10-Day Volume*$_{i,t}$ – The average of daily share volume (in million) over the past 10 trading days. Source: CRSP.
- *PE*$_{i,t}$ – Price-to-earnings ratio, measured for the fiscal year prior to week *t*. Source: Compustat
- *Book Equity*$_{i,t}$ – Log of the book value of equity (in million), measured for the fiscal year prior to week *t*. Source: Compustat.

- *Sale*$_{i,t}$ – Log of sales (saleq, in million), measured for the fiscal year prior to week *t*. Source: Compustat.
- *EBITDA*$_{i,t}$ – Earnings before interest, taxes, depreciation, and amortization, measured for the fiscal year prior to week *t*. Source: Compustat.
- *EPS*$_{i,t}$ – Earnings per share, measured for the fiscal year prior to week *t*. Source: Compustat.
- *Gross Margin*$_{i,t}$ – Gross profits scaled by sales, measured for the fiscal year prior to week *t*. Source: Compustat.
- *Net Margin*$_{i,t}$ – EBITDA scaled by sales, measured for the fiscal year prior to week *t*. Source: Compustat.
- *ROE*$_{i,t}$ – Return on equity, measured by net income divided by equity for the fiscal year prior to week *t*. Source: Compustat.
- *Book-to-Market*$_{i,t}$ – Log of the book value of equity divided by the market value of equity, measured for the fiscal year prior to week *t*. Source: Compustat.
- *Debt-to-Equity*$_{i,t}$ – Log of the book value of total debt divided by the market value of equity, measured for the fiscal year prior to week *t*. Source: Compustat.

Figure 1. Price Charts for Forcerank Contest Stocks. The plots show an example set of 12-week price charts for Forcerank contest stocks. For each Forcerank contest, we provide a corresponding set of historical price figures to GPT-4o and prompt it to issue performance rankings for the ten contest stocks over the following week.

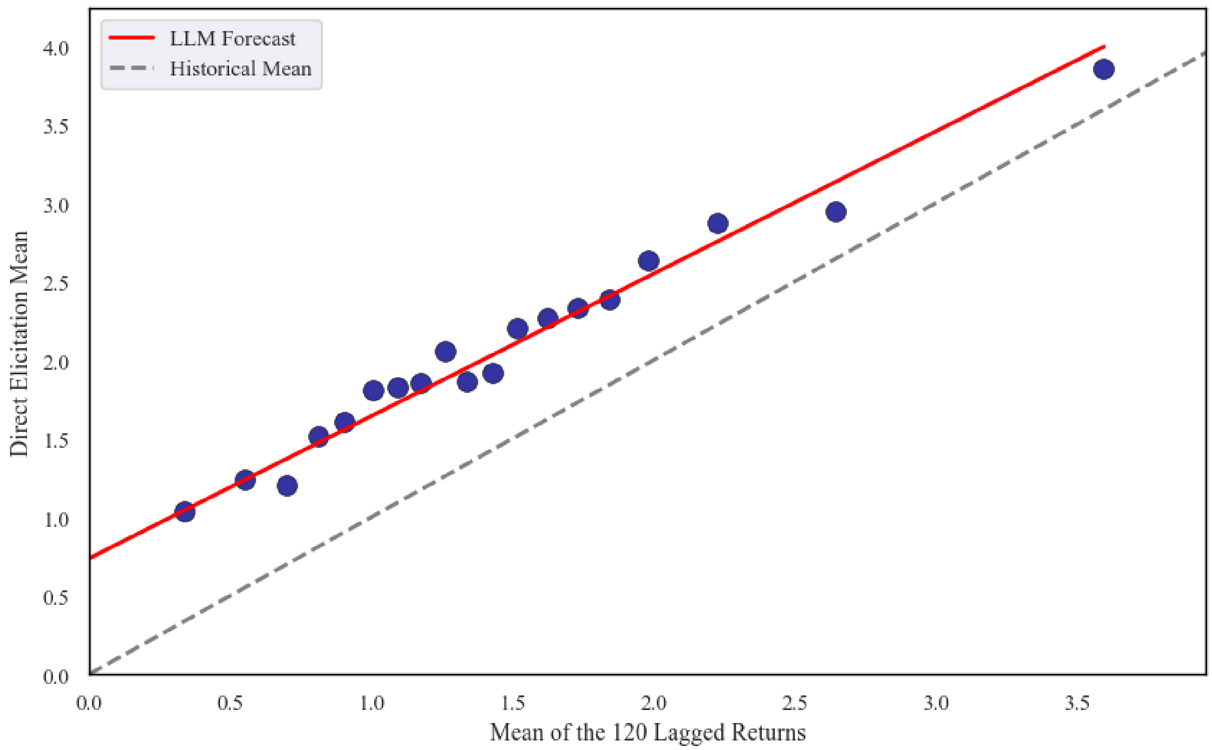

Figure 2. Paired Comparisons of LLM Forecasts and Historical Return Moments. GPT-4o is provided with up to ten (no fewer than five) years of historical monthly returns for a randomly chosen stock-month, and the process is repeated 500 times. The plot shows paired comparisons between LLM-elicited forecasts and the corresponding historical return moments in each prompt, using a scattered bin plot. Blue dots represent the average forecast within bins of the historical value, the red line is a fitted regression line, and the gray dashed line indicates perfect alignment between forecast and historical value.

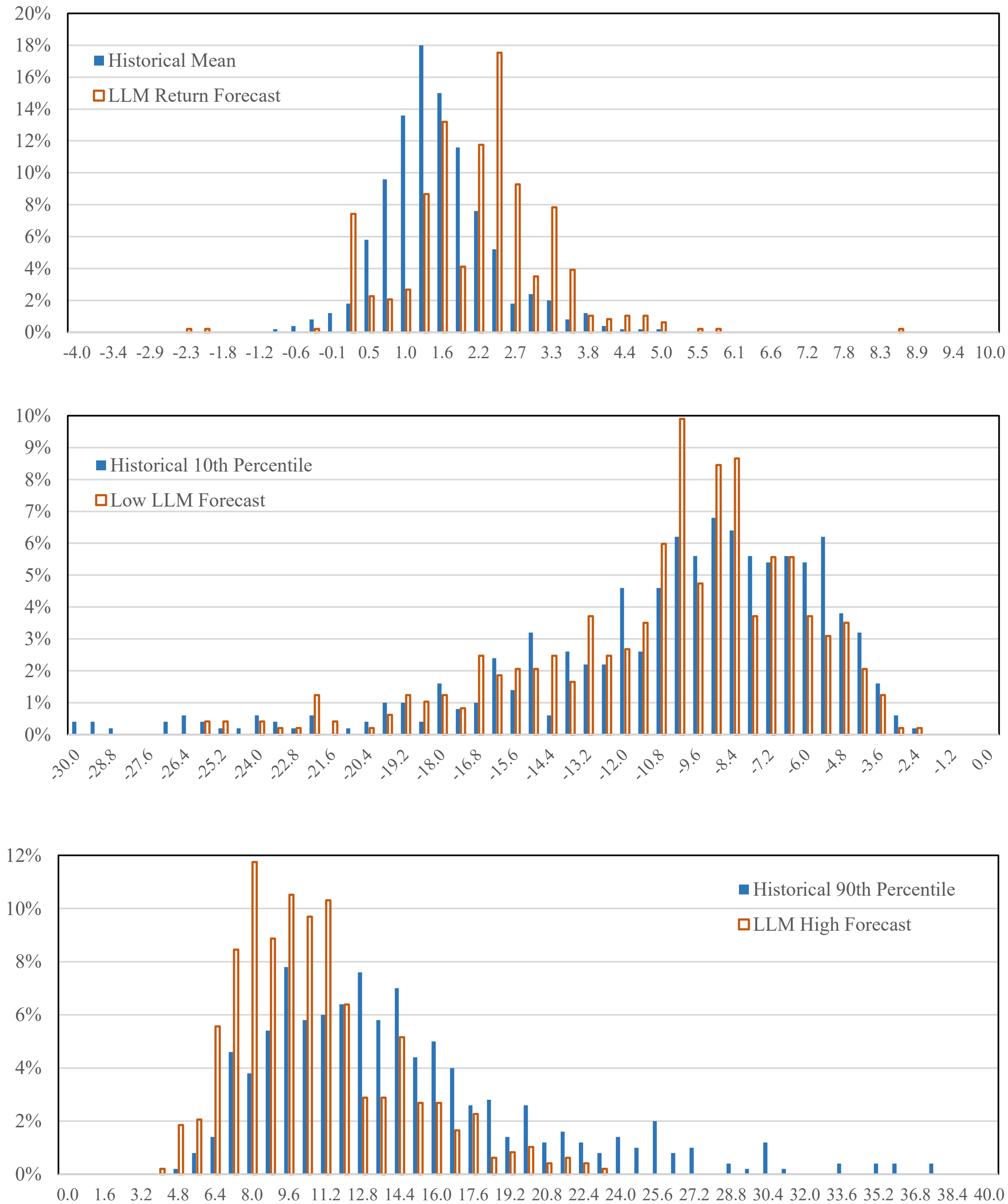

Figure 3. Historical and LLM Forecasts of Low, Expected, and High Returns. GPT-4o is provided with up to ten (no fewer than five) years of historical monthly returns for a randomly chosen stock-month, and the process is repeated 500 times. The top panel plots the distribution of historical means for the data provided in the prompts and the resulting GPT-4o expected return forecasts. The middle plot shows the distribution of the historical 10th percentiles and the low GPT-4o forecast, and the bottom plot shows the 90th percentiles and the high GPT-4o forecasts.

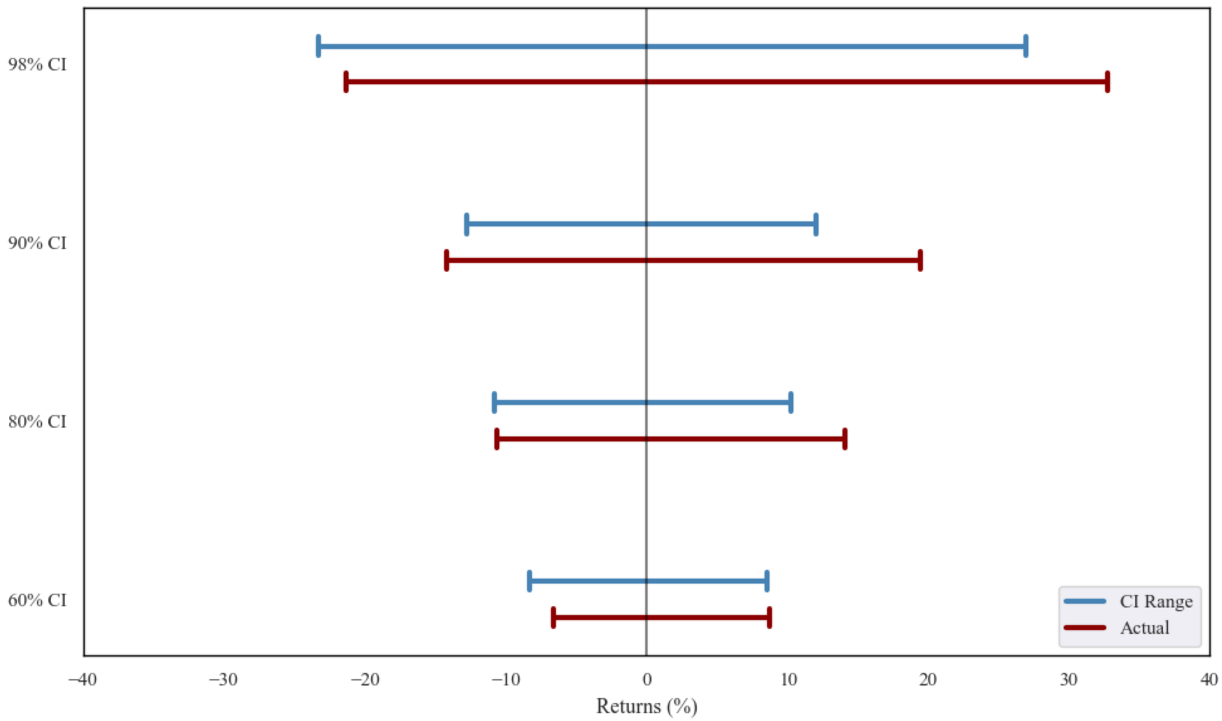

Figure 4. LLM Forecasted Confidence Intervals vs. Historical Return Dispersion. The figure compares LLM-elicited confidence intervals for next-month stock returns to the actual historical return dispersion in the data it was provided. In each of 500 trials, the LLM is prompted to report different symmetric confidence intervals based on up to 10 years (minimum 5 years) of monthly return data for randomly selected stock-months. Blue bars represent the average width of the LLM's predicted intervals (High – Low forecast) for each confidence level. Red bars depict the corresponding empirical intervals computed from the historical data used in the prompt.

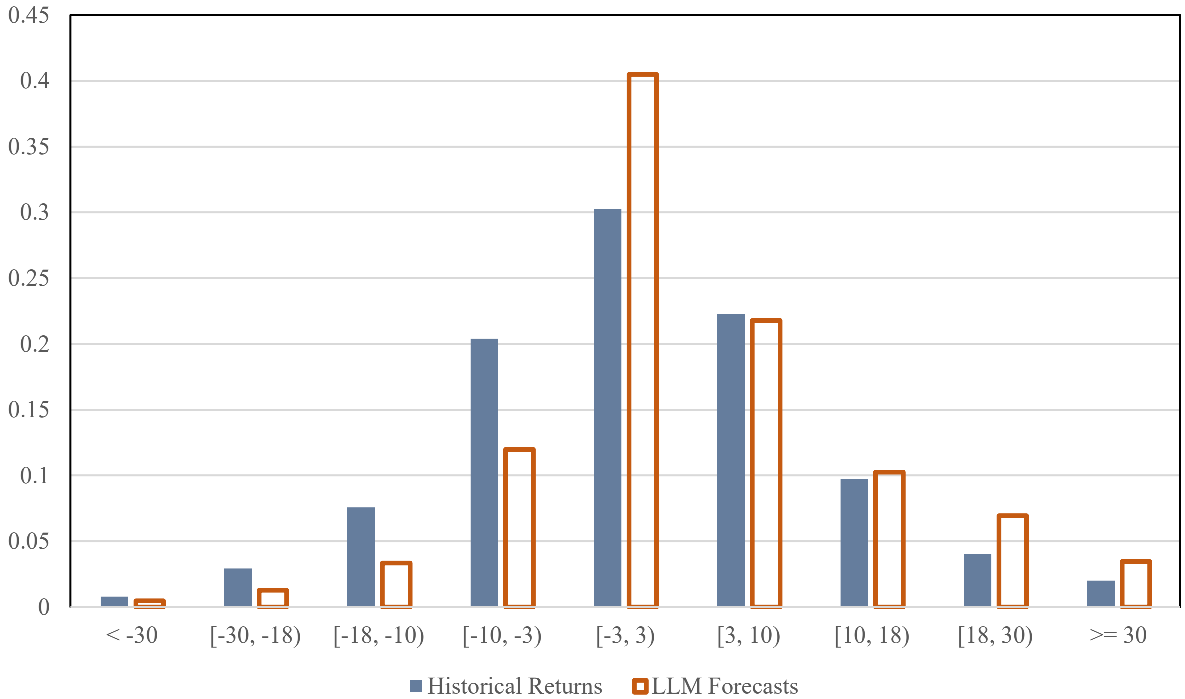

Figure 5. Distribution Histogram of LLM Responses to Stock Return Data. The figure compares the return distributions implied by GPT-4o's forecasts to the actual historical return distributions provided in the prompts. In each of 500 trials, the model receives a randomly selected stock-month along with up to ten years (minimum five years) of historical monthly returns. GPT-4o is asked to assign probabilities that next month's return will fall into each of several predefined return bins. The outlined bars show the average forecasted likelihood across the 500 trials, while the solid bars depict the corresponding empirical frequencies from the historical return data.

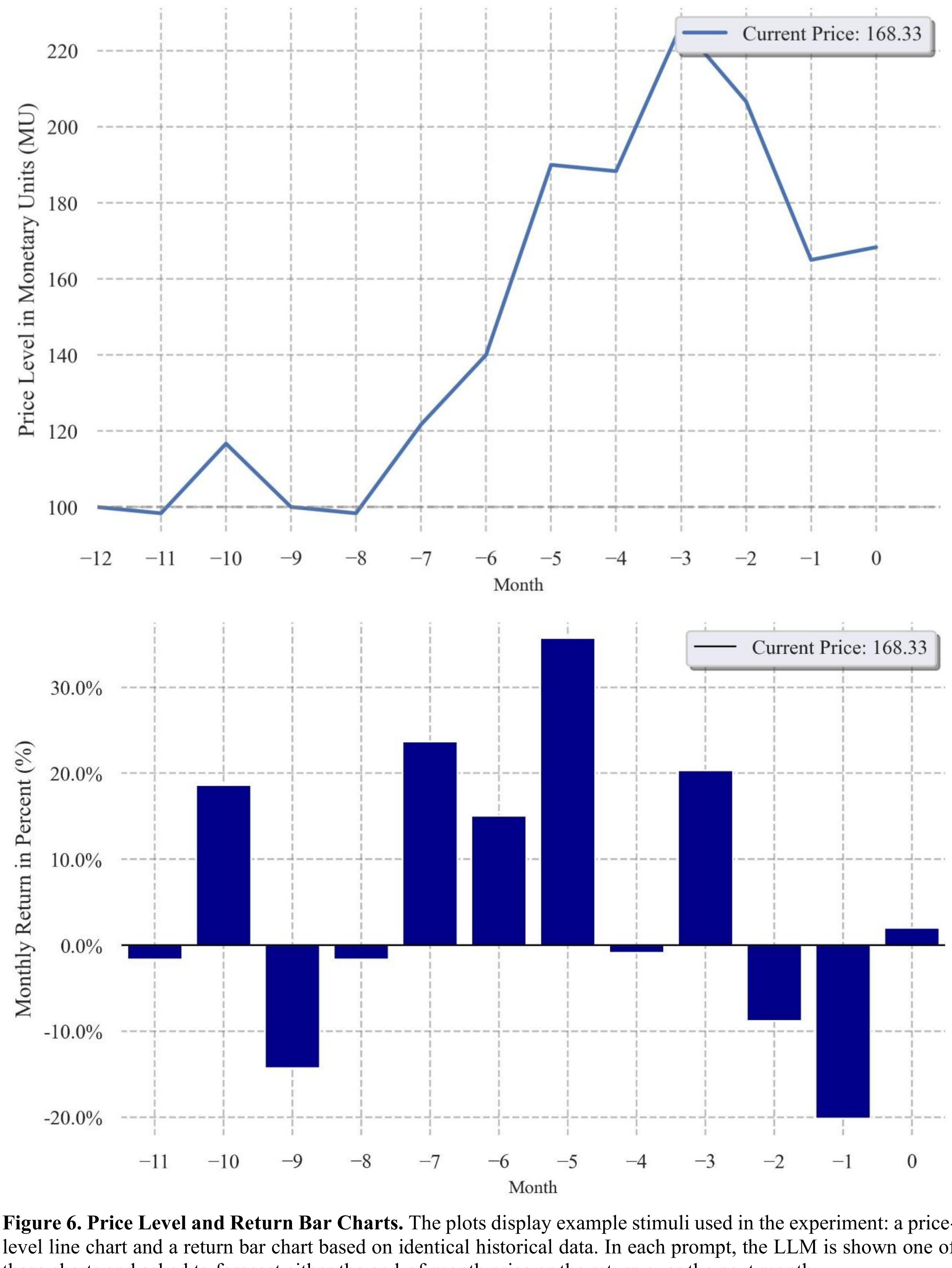

Figure 6. Price Level and Return Bar Charts. The plots display example stimuli used in the experiment: a price-level line chart and a return bar chart based on identical historical data. In each prompt, the LLM is shown one of these charts and asked to forecast either the end-of-month price or the return over the next month.

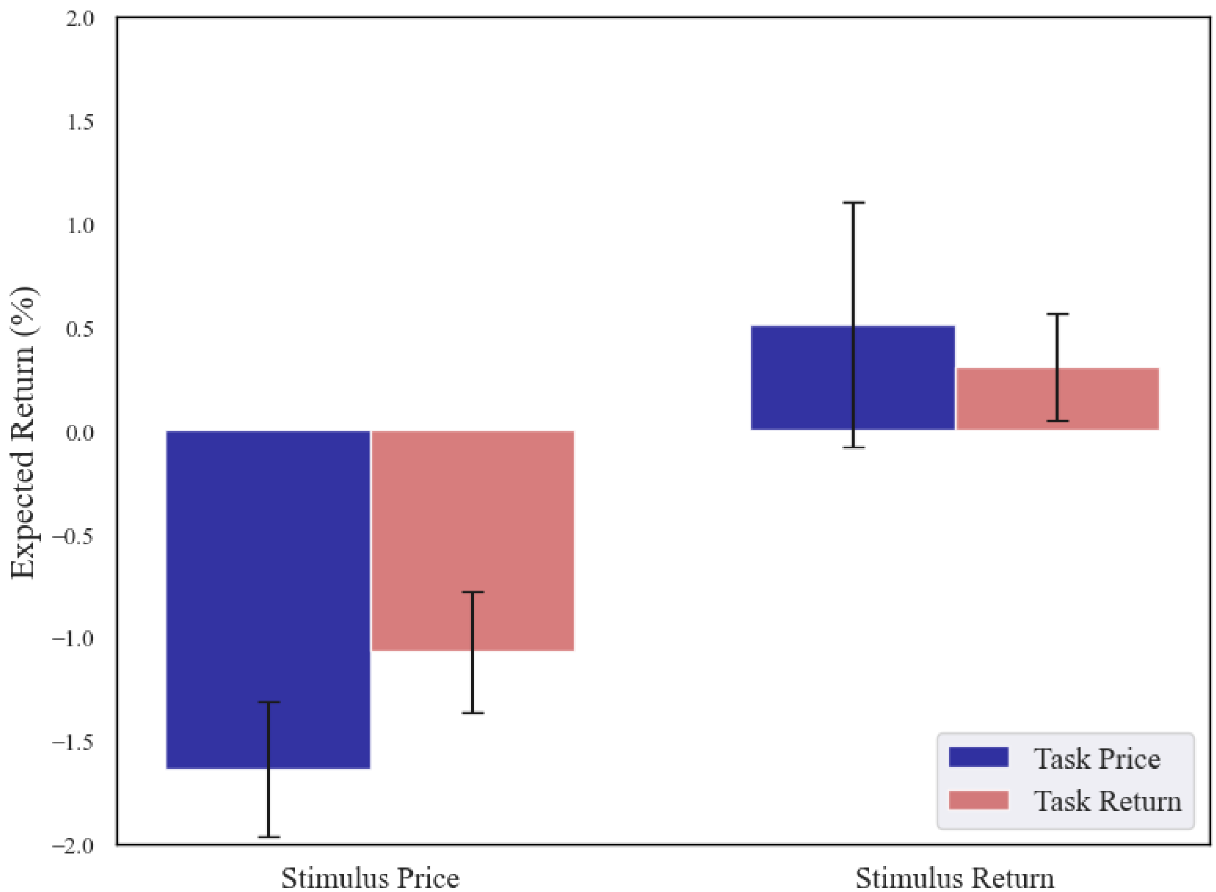

Figure 7. LLM Return Forecasts: Effects of Task Framing and Information Presentation. The figure presents the average excess return forecasts generated by the LLM across 500 stock-month observations, with 95% confidence intervals indicated for each of the four prompt conditions. For prompts requesting a price-level forecast, we convert the predicted price into an implied return using the most recent observed price from the input chart. Excess returns are calculated by subtracting the mean of the historical return data embedded in the chart provided to the LLM.

Table 1. Sample Statistics

The table presents sample descriptive statistics. Panel A provides contest-stock level statistics for the sample of 1,283 Forcerank contests that occurred between February 2016 and December 2017. Panel B provides statistics for monthly observations from the American Association of Individual Investors survey sample, covering July 1987 through June 2024. In each panel, we also include statistics for GPT-4o produced performance rank and sentiment forecasts. Detailed definitions can be found in the Appendix.

Panel A: Contest-Stock level sample

	Obs.	Mean	Standard Deviation	25th Percentile	Median	75th Percentile
Realized Returns (%)	12,719	0.43	4.25	-1.61	0.30	2.38
Forcerank Score	12,719	5.54	2.86	3.00	6.00	8.00
GPT-4o-12week	12,719	5.54	2.85	3.00	6.00	8.00
GPT-4o-24week	12,719	5.54	2.85	3.00	6.00	8.00
Market Capitalization	12,719	9.75	1.66	8.36	9.76	10.98
Book to Market	10,872	-1.46	0.93	-1.95	-1.40	-0.85
Asset Growth	10,872	0.12	0.35	-0.02	0.06	0.17
Profitability	10,872	0.35	0.23	0.20	0.33	0.46
Market Beta	10,872	1.26	0.93	0.71	1.15	1.68
Return MAX	10,872	0.04	0.03	0.02	0.03	0.04
Turnover	10,872	0.48	2.44	0.03	0.05	0.08

Panel B: Monthly sample

	Obs.	Mean	Standard Deviation	25th Percentile	Median	75th Percentile
S&P 500 Returns (%)	438	0.73	4.40	-1.78	1.17	3.52
AAII Bull – Bear	438	0.06	0.18	-0.07	0.06	0.20
GPT-4o Sentiment	438	0.37	0.72	0.00	0.00	1.00

Table 2. Extrapolation of Past Returns: Humans, GPT-4o, and Realized Returns

This table presents the results from linear regressions at the contest-stock-week level, as specified in Eq. (1) in the main text. Specification (1) uses the consensus Forcerank ranking (ranging from one to ten) as the dependent variable, representing the average ranking of a stock across all participants in a contest, with ten indicating the highest rank and one the lowest. In specifications (2) to (4), the dependent variable is the stock ranking generated by GPT-4o based on stock returns in the past 12 weeks. Specification (5) focuses on one-week-ahead stock returns as the dependent variable. The explanatory variables include lagged returns from week $t-11$ to week t. Standard errors are clustered by contest, and *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample period is from February 2016 to December 2017.

	$\text{Forcerank}_{i,t+1}$		$\text{GPT}-\text{4o Rank}^{12w}_{i,t+1}$		$\text{Return}_{i,t+1}$
	(1)	(2)	(3)	(4)	(5)
$\text{Forcerank}_{i,t}$			0.28***	0.17***	
			(27.53)	(21.33)	
$\text{Return}_{i,t}$	12.62***	38.84***		36.66***	1.15
	(20.12)	(38.73)		(37.64)	(0.79)
$\text{Return}_{i,t-1}$	2.31***	2.94***		2.54***	-3.06**
	(4.36)	(4.91)		(4.43)	(-2.31)
$\text{Return}_{i,t-2}$	2.18***	1.92***		1.54***	-3.90***
	(4.22)	(3.38)		(2.83)	(-2.78)
$\text{Return}_{i,t-3}$	2.58***	0.87		0.43	-2.80*
	(5.00)	(1.57)		(0.80)	(-1.93)
$\text{Return}_{i,t-4}$	2.25***	1.15*		0.76	0.84
	(4.26)	(1.95)		(1.33)	(0.61)
$\text{Return}_{i,t-5}$	2.32***	2.08***		1.68***	3.07**
	(4.46)	(3.72)		(3.14)	(2.26)
$\text{Return}_{i,t-6}$	1.53***	-0.38		-0.64	5.32***
	(3.03)	(-0.65)		(-1.16)	(3.54)
$\text{Return}_{i,t-7}$	1.32**	0.42		0.19	-1.35
	(2.53)	(0.68)		(0.33)	(-0.97)
$\text{Return}_{i,t-8}$	1.22**	0.72		0.51	-1.59
	(2.42)	(1.24)		(0.93)	(-1.08)
$\text{Return}_{i,t-9}$	0.83*	0.64		0.50	-3.04**
	(1.68)	(1.14)		(0.92)	(-2.16)
$\text{Return}_{i,t-10}$	1.57***	1.87***		1.60***	-0.93
	(3.29)	(3.36)		(2.99)	(-0.64)
$\text{Return}_{i,t-11}$	0.23	1.10**		1.06**	-1.79
	(0.49)	(2.09)		(2.09)	(-1.37)
Observations	12,668	12,668	12,719	12,668	12,668
R-squared	0.043	0.353	0.079	0.381	0.011

Table 3. Asymmetric Extrapolation of Positive and Negative Returns.

This table presents the results from linear regressions in which the explanatory variables are the positive and negative components of 12 weekly return lags. The positive component of a stock return is defined as Max(Return,0) and the negative component is defined as Min(Return,0). Specification (1) uses the consensus Forcerank ranking (ranging from one to ten) as the dependent variable, representing the average ranking of a stock across all participants in a contest, with ten indicating the highest rank and one the lowest. In Specification (2), the dependent variable is the stock ranking generated by GPT-4o based on the past 12 weekly returns, and in Specification (3), the stock ranking generated by GPT-4o is based on 24 weekly returns. Standard errors are clustered by contest, and *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample period is from February 2016 to December 2017.

	$\text{Forcerank}_{i,t+1}$		$\text{GPT}-\text{4o Rank}_{i,t+1}^{12w}$		$\text{GPT}-\text{4o Rank}_{i,t+1}^{24w}$	
	Positive Returns	Negative Returns	Positive Returns	Negative Returns	Positive Returns	Negative Returns
	(1)	(2)	(3)	(4)	(5)	(6)
$\text{Return}_{i,t}$	9.50***	16.71***	41.11***	36.07***	39.23***	26.30***
	(9.77)	(13.31)	(23.66)	(21.09)	(24.16)	(17.68)
$\text{Return}_{i,t-1}$	-0.01	5.12***	4.19***	1.26	15.80***	1.68*
	(-0.02)	(4.78)	(4.36)	(1.31)	(13.41)	(1.81)
$\text{Return}_{i,t-2}$	-1.03	6.28***	0.95	3.17***	2.38***	2.47**
	(-1.20)	(6.71)	(1.17)	(3.00)	(2.90)	(2.32)
$\text{Return}_{i,t-3}$	0.57	5.21***	0.32	1.64*	1.64**	2.69***
	(0.64)	(5.33)	(0.39)	(1.69)	(2.02)	(2.75)
$\text{Return}_{i,t-4}$	-0.06	5.81***	0.44	2.17**	5.41***	2.87***
	(-0.07)	(5.15)	(0.50)	(2.07)	(5.22)	(2.76)
$\text{Return}_{i,t-5}$	-1.74*	6.72***	1.00	3.55***	5.09***	4.04***
	(-1.94)	(6.90)	(1.20)	(3.52)	(5.35)	(4.11)
$\text{Return}_{i,t-6}$	-0.96	4.13***	-0.05	-0.49	2.21***	0.02
	(-1.15)	(4.02)	(-0.06)	(-0.45)	(2.73)	(0.01)
$\text{Return}_{i,t-7}$	-1.65**	4.89***	1.64*	-1.00	1.86**	0.92
	(-2.00)	(4.69)	(1.90)	(-0.87)	(2.22)	(0.80)
$\text{Return}_{i,t-8}$	-1.49*	4.40***	-0.72	2.52**	-1.01	3.90***
	(-1.75)	(4.53)	(-0.82)	(2.32)	(-1.18)	(3.83)
$\text{Return}_{i,t-9}$	-1.40*	3.14***	-0.37	1.89**	-0.31	1.79*
	(-1.71)	(3.09)	(-0.42)	(1.97)	(-0.37)	(1.95)
$\text{Return}_{i,t-10}$	-0.70	3.73***	1.11	2.67***	0.59	3.10***
	(-0.85)	(3.49)	(1.26)	(2.86)	(0.69)	(2.98)
$\text{Return}_{i,t-11}$	-1.57*	1.89*	1.34	0.66	0.56	1.30
	(-1.93)	(1.91)	(1.61)	(0.78)	(0.67)	(1.50)
Observations	12,719		12,719		12,719	
R-squared	0.073		0.356		0.305	

Table 4. Extrapolation of Past Return Ranks

This table repeats the regression analysis in Table 2, employing return ranks as the explanatory variables (i.e., the stocks' actual past rankings converted from past weekly returns in the contest). Specification (1) uses the consensus Forcerank ranking (ranging from one to ten) as the dependent variable, representing the average ranking of a stock across all participants in a contest, with ten indicating the highest rank and one the lowest. In specifications (2) to (4), the dependent variable is the stock ranking generated by GPT-4o based on unadjusted stock returns in the past 12 weeks. Specification (5) focuses on the realized one-week-ahead stock ranks as the dependent variable. The explanatory variables include lagged returns from week $t-11$ to week t. Standard errors are clustered by contest, and *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample period is from February 2016 to December 2017.

	$\text{Forcerank}_{i,t+1}$		$\text{GPT}-\text{4o Rank}_{i,t+1}^{12w}$		$\text{Ret Rank}_{i,t+1}$
	(1)	(2)	(3)	(4)	(5)
$\text{Forcerank}_{i,t}$			0.28***	0.00	
			(27.53)	(1.01)	
$\text{Return Rank}_{i,t}$	0.30***	0.91***		0.91***	-0.02
	(29.26)	(222.18)		(215.74)	(-1.49)
$\text{Return Rank}_{i,t-1}$	0.04***	0.07***		0.07***	-0.01
	(4.81)	(16.30)		(16.26)	(-1.28)
$\text{Return Rank}_{i,t-2}$	0.05***	0.02***		0.02***	-0.02*
	(5.06)	(4.40)		(4.33)	(-1.79)
$\text{Return Rank}_{i,t-3}$	0.05***	0.01***		0.01**	-0.00
	(5.34)	(2.61)		(2.55)	(-0.24)
$\text{Return Rank}_{i,t-4}$	0.04***	0.01***		0.01***	0.02**
	(4.68)	(3.56)		(3.51)	(2.43)
$\text{Return Rank}_{i,t-5}$	0.04***	0.01***		0.01***	0.02
	(3.86)	(3.37)		(3.33)	(1.50)
$\text{Return Rank}_{i,t-6}$	0.03***	0.01**		0.01**	0.03***
	(3.44)	(2.38)		(2.34)	(3.08)
$\text{Return Rank}_{i,t-7}$	0.03***	0.01***		0.01***	0.00
	(3.37)	(2.84)		(2.80)	(0.24)
$\text{Return Rank}_{i,t-8}$	0.02***	0.01		0.01	0.00
	(2.62)	(1.52)		(1.49)	(0.17)
$\text{Return Rank}_{i,t-9}$	0.02**	0.01**		0.01**	0.02*
	(2.31)	(2.18)		(2.16)	(1.72)
$\text{Return Rank}_{i,t-10}$	0.02**	0.01**		0.01**	-0.01
	(2.41)	(2.37)		(2.34)	(-0.96)
$\text{Return Rank}_{i,t-11}$	0.01	0.02***		0.02***	-0.01
	(1.22)	(4.70)		(4.69)	(-0.86)
Observations	12,719	12,719	12,719	12,719	12,719
R-squared	0.102	0.836	0.079	0.836	0.003

Table 5. Extrapolative Beliefs: Exponential Decay Model.

The table presents the results of a contest-level nonlinear regression specified in Eq. (3) of the main text:

$$Y_{i,t+1} = 5.5 + \lambda_1 \cdot \sum_{s=0}^{12} w_s R_{i,t-s} + \epsilon_{i,t}, \quad \text{where } w_s = \frac{\lambda_2^s}{\sum_{j=0}^{12} \lambda_2^j}$$

In Specification (1), the dependent variable is the consensus ranking (one to ten) representing a stock's average ranking across all contest participants. In Specification (2), the dependent variable is the ranking produced by GPT-4o using 12 weekly return lags, and in Specification (3), 24 return lags are considered. The explanatory variables include lagged returns from week $t-11$ to week *t*. The exponential decay model is estimated using GMM, and *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. DHJ show theoretically that a higher λ_1 and a lower λ_2 jointly lead to a higher degree of extrapolation and $\lambda_1(1-\lambda_2)$ represents the degree of extrapolation. The sample period is from February 2016 to December 2017.

	$\text{Forcerank}_{i,t}$	$\text{GPT}-\text{4o Rank}_{i,t}^{12w}$	$\text{GPT}-\text{4o Rank}_{i,t}^{24w}$
	(1)	(2)	(3)
λ_1	16.98***	40.72***	45.68***
	(15.53)	(48.89)	(45.30)
λ_2	0.28***	0.07***	0.27***
	(6.78)	(5.16)	(18.55)
$\lambda_1(1-\lambda_2)$	12.19	38.03	33.21

Table 6. GPT-4o Forecast Ranks and Future Stock Returns

This table presents the results from Fama-MacBeth return forecasting regressions. For each week *t* and each stock *i*, the dependent variable is the daily return of stock *i* over week *t*+1. The return predictors include the GPT-4o stock rank and its decomposed components: the predicted component is derived as the fitted value from the nonlinear regression specified in Equation (3), while the residual component is referred to as the residual GPT-4o rank. Panel A focuses on the ranking produced by GPT-4o using 12 weekly return lags, and Panel B focuses on GPT-4o ranks based on 24 return lags. Control variables measured at week *t*, include log market capitalization, log book-to-market, asset growth, gross profits-to-assets, market beta, weekly turnover, and the max daily return in the last month. Returns are measured in basis points, with t-statistics provided in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively. The sample period is from February 2016 to December 2017.

Panel A: Forecast inferred from 12 lagged returns

	(1)	(2)	(3)	(4)	(5)	(6)
$\text{GPT} - \text{4o Rank}_{i,t}^{12w}$	-0.12			-0.41		
	(-0.35)			(-1.15)		
$\text{Predicted GPT} - \text{4o Rank}_{i,t}^{12w}$		-0.38			-1.08*	
		(-0.76)			(-1.92)	
$\text{Residual GPT} - \text{4o Rank}_{i,t}^{12w}$			0.11			0.35
			(0.18)			(0.56)
Controls	No	No	No	Yes	Yes	Yes
Observations	58,056	58,056	58,056	49,683	49,683	49,683
R-squared	0.012	0.017	0.013	0.183	0.189	0.184

Panel B: Forecast inferred from 24 lagged returns

	(1)	(2)	(3)	(4)	(5)	(6)
$\text{GPT} - \text{4o Rank}_{i,t}^{24w}$	-0.50			-0.78**		
	(-1.40)			(-2.05)		
$\text{Predicted GPT} - \text{4o Rank}_{i,t}^{24w}$		-0.69			-1.34**	
		(-1.25)			(-2.20)	
$\text{Residual GPT} - \text{4o Rank}_{i,t}^{24w}$			-0.67			-0.67
			(-1.35)			(-1.18)
Controls	No	No	No	Yes	Yes	Yes
Observations	58,056	58,056	58,056	49,683	49,683	49,683
R-squared	0.013	0.017	0.012	0.184	0.188	0.185

Table 7. LLM Extrapolation of Price Charts

This table presents the results from linear regressions at the contest-stock-week level, as specified in Eq. (1) in the main text. Forcerank is the average ranking of a stock across all participants in a contest, with ten indicating the highest rank and one the lowest. The dependent variable is the performance ranking generated by GPT-4o based on price charts over the past 12 weeks. The explanatory variables include lagged returns relative to forecast week *t*+1. Standard errors are clustered by contest, and *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample period is from February 2016 to December 2017.

	$\text{GPT}-4\text{o Rank}_{i,t+1}^{12w\ Chart}$		
	(1)	(2)	(3)
$\text{Forcerank}_{i,t}$		0.05***	0.02**
		(5.06)	(2.43)
$\text{Return}_{i,t}$	6.29***		6.00***
	(11.62)		(11.11)
$\text{Return}_{i,t-1}$	5.00***		4.94***
	(9.67)		(9.57)
$\text{Return}_{i,t-2}$	3.73***		3.68***
	(6.93)		(6.84)
$\text{Return}_{i,t-3}$	2.79***		2.73***
	(5.21)		(5.11)
$\text{Return}_{i,t-4}$	2.33***		2.28***
	(4.60)		(4.51)
$\text{Return}_{i,t-5}$	2.07***		2.01***
	(3.93)		(3.84)
$\text{Return}_{i,t-6}$	2.45***		2.41***
	(4.71)		(4.64)
$\text{Return}_{i,t-7}$	1.81***		1.78***
	(3.47)		(3.42)
$\text{Return}_{i,t-8}$	1.23**		1.20**
	(2.53)		(2.48)
$\text{Return}_{i,t-9}$	0.63		0.61
	(1.30)		(1.26)
$\text{Return}_{i,t-10}$	1.81***		1.77***
	(3.57)		(3.51)
$\text{Return}_{i,t-11}$	0.79*		0.78*
	(1.69)		(1.68)
Observations	12,668	12,719	12,668
R-squared	0.024	0.002	0.025

Table 8. Market Return Extrapolation: AAII and GPT-4o Sentiment

This table presents the results from linear regressions of sentiment on lagged aggregate returns. In Specification (1), the dependent variable is AAII sentiment, measured as the percentage of "bullish" investors minus the percentage of "bearish" investors in the last week of each month. In Specifications (2) to (4), the dependent variable is the GPT-4o sentiment generated based on US stock market (S&P 500) returns in the past 12 months. In Specification (5), the dependent variable is the cumulative realized S&P 500 return over the next six months. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample period is from July 1987 to June 2024.

	AAII Sent_{t+1}	$\text{GPT}-\text{4o Sentiment}_{t+1}$			$\text{S\&P Ret}_{t+1,t+6}$
	(1)	(2)	(3)	(4)	(5)
AAII Sent_t			1.63***	0.30**	
			(8.92)	(2.57)	
$\text{S\&P Return}_t$	1.42***	9.17***		8.74***	0.01
	(6.60)	(16.68)		(15.14)	(0.05)
$\text{S\&P Return}_{t-1}$	0.80***	5.38***		5.14***	0.13
	(3.99)	(10.20)		(9.38)	(0.95)
$\text{S\&P Return}_{t-2}$	0.38**	3.99***		3.87***	0.05
	(1.99)	(8.78)		(8.41)	(0.38)
$\text{S\&P Return}_{t-3}$	0.35*	3.65***		3.55***	0.09
	(1.77)	(7.48)		(7.26)	(0.65)
$\text{S\&P Return}_{t-4}$	0.26	2.68***		2.60***	0.04
	(1.35)	(4.89)		(4.76)	(0.30)
$\text{S\&P Return}_{t-5}$	0.22	3.39***		3.32***	0.11
	(1.14)	(7.04)		(6.79)	(0.84)
$\text{S\&P Return}_{t-6}$	0.12	2.48***		2.44***	-0.05
	(0.60)	(6.03)		(5.88)	(-0.38)
$\text{S\&P Return}_{t-7}$	0.18	2.00***		1.94***	-0.13
	(0.93)	(4.43)		(4.35)	(-1.07)
$\text{S\&P Return}_{t-8}$	0.33	1.78***		1.68***	-0.05
	(1.63)	(3.69)		(3.53)	(-0.41)
$\text{S\&P Return}_{t-9}$	0.10	2.13***		2.10***	-0.06
	(0.50)	(4.83)		(4.81)	(-0.55)
$\text{S\&P Return}_{t-10}$	0.25	1.47***		1.39***	0.00
	(1.40)	(3.26)		(3.08)	(0.01)
$\text{S\&P Return}_{t-11}$	0.04	2.12***		2.11***	-0.06
	(0.21)	(5.11)		(5.04)	(-0.55)
Observations	438	438	438	438	438
R-squared	0.194	0.687	0.168	0.692	0.012

Table 9. Market Return Extrapolation: LLM Prompt Engineering

This table presents the results from linear regressions at the month level. The dependent variable is the GPT-4o sentiment measure at month *t* when provided with 12 lagged monthly returns for the S&P 500 index and asked, "Do you feel the direction of the stock market over the next six months will be up (bullish), no change (neutral), or down (bearish)?" A score of 1 represents bullish sentiment, 0 represents neutral sentiment, and -1 represents bearish sentiment. In Specification (1), the prompt is followed by "Think step by step when creating your response." In Specification (2), the prompt is followed by "Analyze the data and consider an appropriate model to apply for your response." In Specification (3), the prompt is followed by "Consider human biases that may affect responses to similar questions and avoid these biases when creating your response." Specification 4 uses a revised prompt based on inferences from Greenwood and Shleifer (2014). Specification (5) presents the baseline as in Specification (2) of Table 8. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample period is from July 1987 to June 2024.

	$\text{GPT}-4\text{o Sentiment}_{t+1}$				
	(1)	(2)	(3)	(4)	(5)
$\text{S\&P Return}_t$	9.06***	7.61***	9.65***	6.21***	9.17***
	(16.10)	(11.98)	(17.10)	(12.37)	(16.68)
$\text{S\&P Return}_{t-1}$	6.41***	6.58***	5.27***	4.43***	5.38***
	(10.65)	(9.81)	(8.78)	(8.79)	(10.20)
$\text{S\&P Return}_{t-2}$	4.79***	4.42***	3.36***	3.59***	3.99***
	(8.51)	(6.53)	(6.18)	(7.88)	(8.78)
$\text{S\&P Return}_{t-3}$	3.20***	3.45***	2.67***	2.73***	3.65***
	(5.25)	(5.41)	(5.20)	(5.79)	(7.48)
$\text{S\&P Return}_{t-4}$	2.33***	2.56***	1.69***	2.52***	2.68***
	(4.06)	(3.98)	(3.13)	(5.00)	(4.89)
$\text{S\&P Return}_{t-5}$	2.41***	2.59***	1.66***	2.20***	3.39***
	(4.48)	(4.33)	(2.92)	(4.82)	(7.04)
$\text{S\&P Return}_{t-6}$	3.06***	2.22***	1.91***	1.78***	2.48***
	(6.75)	(3.47)	(3.60)	(4.31)	(6.03)
$\text{S\&P Return}_{t-7}$	1.71***	1.88***	0.97**	1.79***	2.00***
	(3.59)	(2.85)	(2.15)	(4.11)	(4.43)
$\text{S\&P Return}_{t-8}$	1.59***	1.44**	2.25***	0.98**	1.78***
	(3.21)	(2.42)	(4.67)	(2.23)	(3.69)
$\text{S\&P Return}_{t-9}$	1.44***	2.53***	1.01**	1.28***	2.13***
	(2.95)	(3.84)	(2.08)	(3.18)	(4.83)
$\text{S\&P Return}_{t-10}$	0.34	1.71***	1.31***	1.41***	1.47***
	(0.73)	(2.93)	(2.89)	(3.36)	(3.26)
$\text{S\&P Return}_{t-11}$	0.85*	0.64	0.97*	1.75***	2.12***
	(1.79)	(1.25)	(1.95)	(4.31)	(5.11)
Observations	438	438	438	438	438
R-squared	0.630	0.521	0.625	0.532	0.687

Table 10. LLM Forecasts of the Return Distribution – Bias Tests and Calibration Evidence

The table presents descriptive statistics and bias tests for the return distribution forecasts generated by GPT-4o. Panel A summarizes historical returns, GPT-4o forecasts, and next-month realized returns. Panel B tests for forecast bias, comparing the expected values against historical means and realized outcomes. Panel C evaluates calibration, focusing on the accuracy of the Low and High forecasts, which are interpreted as the 10th and 90th percentiles and intended to represent an 80% confidence interval for the next month's return. All statistics are based on 500 stock-month observations randomly selected from the 1926 to 2023 period, with up to ten years of historical monthly returns provided to GPT-4o.

Panel A: Descriptive statistics

Variable	Obs.	Mean	Std.	5%	25%	50%	75%	95%
Expected forecast	485	2.00	1.11	0.00	1.34	2.12	2.53	3.56
Historical mean	485	1.38	0.83	0.25	0.87	1.29	1.80	2.97
Realized returns	499	1.12	11.17	-15.66	-5.56	0.00	7.12	19.75
Low forecast	485	-10.83	4.37	-19.32	-12.95	-9.97	-7.78	-4.97
Historical 10%	500	-10.70	5.13	-20.87	-12.78	-9.54	-7.09	-4.57
High forecast	485	10.29	3.57	5.70	7.72	9.68	11.88	17.08
Historical 90%	500	14.11	6.01	6.92	9.83	12.84	16.57	25.76
Confidence interval %	485	21.13	7.63	10.96	15.86	19.42	24.87	35.35
Historical 90% - 10%	500	24.81	10.86	11.68	17.11	22.51	29.74	47.67

Panel B: Forecast bias

Difference	Mean Difference	p-Value
Expected forecast – Historical mean	0.61	0.00
Expected forecast – Realized return	0.88	0.11
Low forecast – Historical 10%	-0.14	0.23
High forecast – Historical 90%	-3.82	0.00

Panel C: Realized returns relative to historical and GPT-4o forecasts

% of realized returns below low forecast	11.4
% of realized returns in confidence interval	69.3
% of realized returns above high forecast	19.3
% of realized returns below historical 10%	11.4
% of realized returns in historical interval	78.0
% of realized returns above historical 90%	10.6

Table 11. LLM Forecast Bias Tests and Calibration Evidence: Prompt Engineering

The table presents descriptive statistics and bias tests for the return distribution forecasts generated by GPT-4o. Panel A tests for forecast bias, comparing the expected values against historical means and realized outcomes, with *p*-values from tests of equality of means reported beneath each difference. Panel B evaluates calibration, focusing on the accuracy of the Low and High forecasts, which are interpreted as the 10th and 90th percentiles and intended to represent an 80% confidence interval for the next month's return. All statistics are based on 500 stock-month observations randomly selected from the 1926 to 2023 period, with up to ten years of historical monthly returns provided to GPT-4o. In Specification (1), the prompt is followed by "Think step by step when creating your response." In Specification (2), the prompt is followed by "Analyze the data and consider an appropriate model to apply for your response." In Specification (3), the prompt is followed by "Consider human biases that may affect responses to similar questions and avoid these biases when creating your response." Specification 4 uses a revised prompt based on inferences from Hartzmark and Sussman (2024). Specification (5) presents the baseline evidence as in Table 10.

Panel A: Forecast bias

	Difference between LLM forecast and Historical Data (%)				
Difference	(1)	(2)	(3)	(4)	(5)
Expected forecast – Historical mean	0.56	0.87	0.58	0.74	0.61
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Expected forecast – Realized return	0.83	1.13	0.84	1.01	0.88
	(0.07)	(0.02)	(0.09)	(0.04)	(0.11)
Low forecast – Historical 10%	-0.70	0.17	-0.04	0.74	-0.14
	(0.00)	(0.01)	(0.55)	(0.00)	(0.23)
High forecast – Historical 90%	-3.70	-3.62	-3.44	-3.20	-3.82
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

Panel B: Realized returns relative to historical and LLM forecasts

	(1)	(2)	(3)	(4)	(5)
% of returns < low LLM forecast	10.1	11.7	11.4	11.9	11.4
% of returns in 80% LLM CI	71.9	69.8	70.5	70.7	69.3
% of returns > high LLM forecast	18.0	18.5	18.1	17.5	19.3

Table IA1. Extrapolation of Past Returns: Contest-Adjusted Returns

This table repeats the regression analysis in Table 2, focusing on contest-adjusted returns (i.e., the stock return in excess of the average return of the ten stocks in the contest). Specification (1) uses the consensus Forcerank ranking (ranging from one to ten) as the dependent variable, representing the average ranking of a stock across all participants in a contest, with ten indicating the highest rank and one the lowest. In specifications (2) to (4), the dependent variable is the stock ranking generated by GPT-4o based on unadjusted stock returns in the past 12 weeks. Specification (5) focuses on one-week-ahead stock contest-adjusted returns as the dependent variable. The explanatory variables include lagged returns from week *t* − 11 to week *t*. Standard errors are clustered by contest, and *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

	$\text{Forcerank}_{i,t+1}$	$\text{GPT}-\text{4o Rank}^{12w}_{i,t+1}$			$\text{Adj Return}_{i,t+1}$
	(1)	(2)	(3)	(4)	(5)
$\text{Forcerank}_{i,t}$			0.28***	0.10***	
			(27.20)	(12.82)	
$\text{Adj Return}_{i,t}$	20.17***	62.52***		60.51***	0.01
	(19.64)	(37.44)		(36.10)	(0.75)
$\text{Adj Return}_{i,t-1}$	3.54***	4.38***		4.03***	-0.01
	(4.36)	(6.68)		(6.23)	(-1.32)
$\text{Adj Return}_{i,t-2}$	3.08***	1.72***		1.41***	-0.02**
	(3.90)	(3.36)		(2.76)	(-2.14)
$\text{Adj Return}_{i,t-3}$	4.02***	0.84		0.44	-0.02*
	(5.07)	(1.51)		(0.79)	(-1.67)
$\text{Adj Return}_{i,t-4}$	3.60***	2.07***		1.71***	0.03**
	(4.36)	(3.86)		(3.22)	(2.49)
$\text{Adj Return}_{i,t-5}$	3.08***	1.68***		1.37**	0.01
	(3.84)	(2.97)		(2.45)	(0.88)
$\text{Adj Return}_{i,t-6}$	3.05***	0.90		0.60	0.03**
	(3.88)	(1.62)		(1.10)	(2.32)
$\text{Adj Return}_{i,t-7}$	2.77***	1.75***		1.47**	-0.02*
	(3.38)	(2.73)		(2.35)	(-1.82)
$\text{Adj Return}_{i,t-8}$	2.54***	1.69***		1.44***	0.00
	(3.31)	(3.23)		(2.78)	(0.10)
$\text{Adj Return}_{i,t-9}$	1.57**	0.73		0.57	-0.00
	(1.99)	(1.32)		(1.05)	(-0.20)
$\text{Adj Return}_{i,t-10}$	2.08***	0.80		0.60	-0.01
	(2.67)	(1.43)		(1.07)	(-0.72)
$\text{Adj Return}_{i,t-11}$	0.61	1.60***		1.54***	-0.02
	(0.80)	(2.85)		(2.78)	(-1.47)
Observations	12,752	12,752	12,807	12,752	12,752
R-squared	0.070	0.566	0.078	0.575	0.004

Table IA2. Extrapolation of Past Returns: Humans, GPT-4o, and Realized Returns – 24 Weeks

This table repeats the regression analysis in Table 2, extending the number of weekly return lags from 12 to 24. Specification (1) uses the consensus Forcerank ranking (ranging from one to ten) as the dependent variable, representing the average ranking of a stock across all participants in a contest, with ten indicating the highest rank and one the lowest. In specifications (2) to (4), the dependent variable is the stock ranking generated by GPT-4o based on stock returns in the past 24 weeks. Specification (5) focuses on one-week-ahead stock returns as the dependent variable. The explanatory variables include lagged returns from week $t-23$ to week t. Standard errors are clustered by contest, and *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

	$Forcerank_{i,t+1}$		$GPT-4o^{24w}_{i,t+1}$		$Return_{i,t+1}$
	(1)	(2)	(3)	(4)	(5)
$Forcerank_{i,t}$			0.23***	0.13***	
			(23.63)	(16.19)	
$Return_{i,t}$	12.73***	33.69***		32.04***	1.05
	(20.26)	(37.70)		(36.76)	(0.73)
$Return_{i,t-1}$	2.39***	9.66***		9.35***	-2.67**
	(4.47)	(14.75)		(14.49)	(-2.03)
$Return_{i,t-2}$	2.22***	2.53***		2.25***	-3.79***
	(4.25)	(4.58)		(4.20)	(-2.69)
$Return_{i,t-3}$	2.52***	2.07***		1.74***	-2.64*
	(4.74)	(3.70)		(3.22)	(-1.87)
$Return_{i,t-4}$	2.31***	4.34***		4.04***	0.64
	(4.32)	(6.99)		(6.67)	(0.46)
$Return_{i,t-5}$	2.35***	4.41***		4.11***	2.69**
	(4.44)	(7.65)		(7.32)	(1.99)
$Return_{i,t-6}$	1.59***	0.59		0.39	5.45***
	(3.09)	(1.03)		(0.70)	(3.63)
$Return_{i,t-7}$	1.41***	1.06*		0.88	-1.21
	(2.69)	(1.78)		(1.52)	(-0.88)
$Return_{i,t-8}$	1.32***	1.10**		0.93*	-2.11
	(2.58)	(1.97)		(1.73)	(-1.50)
$Return_{i,t-9}$	0.99**	0.60		0.47	-3.02**
	(1.98)	(1.08)		(0.87)	(-2.17)
$Return_{i,t-10}$	1.83***	1.74***		1.50***	-0.61
	(3.71)	(3.07)		(2.72)	(-0.43)
$Return_{i,t-11}$	0.20	0.89*		0.86*	-2.00
	(0.43)	(1.69)		(1.68)	(-1.51)
$Return_{i,t-12}$	0.80*	0.61		0.51	-1.08
	(1.67)	(1.17)		(1.00)	(-0.74)
$Return_{i,t-13}$	0.99**	1.05**		0.93*	-0.13
	(2.20)	(2.00)		(1.80)	(-0.09)
$Return_{i,t-14}$	1.04**	0.85		0.71	0.71

	(2.33)	(1.53)		(1.32)	(0.55)
$Return_{i,t-15}$	0.97**	1.59***		1.47***	1.15
	(2.03)	(3.05)		(2.90)	(0.73)
$Return_{i,t-16}$	0.92*	-0.56		-0.68	3.84***
	(1.96)	(-0.95)		(-1.18)	(2.66)
$Return_{i,t-17}$	-0.39	-1.05*		-1.00*	-0.59
	(-0.82)	(-1.89)		(-1.85)	(-0.46)
$Return_{i,t-18}$	0.53	0.56		0.49	-3.54**
	(1.14)	(1.07)		(0.95)	(-2.30)
$Return_{i,t-19}$	1.84***	-0.29		-0.53	-3.20**
	(3.94)	(-0.53)		(-1.00)	(-2.30)
$Return_{i,t-20}$	1.12**	1.55***		1.40***	2.67*
	(2.42)	(2.86)		(2.67)	(1.82)
$Return_{i,t-21}$	0.19	0.01		-0.01	0.02
	(0.38)	(0.02)		(-0.03)	(0.02)
$Return_{i,t-22}$	0.77	0.21		0.11	-0.22
	(1.57)	(0.36)		(0.19)	(-0.15)
$Return_{i,t-23}$	-0.01	0.35		0.35	2.31*
	(-0.02)	(0.61)		(0.64)	(1.72)
Observations	12,607	12,607	12,719	12,607	12,607
R-squared	0.046	0.292	0.054	0.308	0.018

Table IA3. LLM Extrapolation of Returns: Simulated Contests in the Post-Training Period
This table presents the results from linear regressions at the simulated contest-stock-week level, as specified in Eq. (1) in the main text. Simulated contests are created by randomly choosing 10 stocks for each contest, and we form two contests for each of the 11 GIC industries each week. The post-training, out-of-sample period is March to November 2024, and the adjacent in-training-sample period is March to December 2023. As in Table 2, GPT-4o predictions are based on 12 weeks of lagged returns, and the explanatory variables include lagged returns relative to forecast week *t*+1. Standard errors are clustered by contest, and *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

	$\text{GPT}-4\text{o Rank}_{i,t+1}^{12w}$		
	Full Sample (Mar 2023-Nov 2024)	Training Period (In Sample) Mar 2023-Dec 2023	Post-Launch (Out of Sample) Apr 2024-Nov 2024
	(1)	(2)	(3)
$\text{Return}_{i,t}$	24.57***	23.30***	25.62***
	(36.69)	(24.58)	(27.74)
$\text{Return}_{i,t-1}$	4.40***	5.43***	3.54***
	(9.02)	(7.82)	(4.39)
$\text{Return}_{i,t-2}$	2.45***	1.57***	3.27***
	(6.70)	(2.65)	(4.24)
$\text{Return}_{i,t-3}$	2.21***	0.57	3.96***
	(4.62)	(0.87)	(4.62)
$\text{Return}_{i,t-4}$	1.94***	1.83***	2.45***
	(4.42)	(3.02)	(2.98)
$\text{Return}_{i,t-5}$	2.07***	1.70**	2.19***
	(4.57)	(2.51)	(2.66)
$\text{Return}_{i,t-6}$	1.53***	0.47	2.50***
	(3.41)	(0.74)	(3.10)
$\text{Return}_{i,t-7}$	1.95***	0.45	2.04***
	(4.58)	(0.75)	(2.78)
$\text{Return}_{i,t-8}$	1.74***	0.96	2.16***
	(4.04)	(1.43)	(2.87)
$\text{Return}_{i,t-9}$	1.44***	-0.27	2.17***
	(3.46)	(-0.44)	(2.93)
$\text{Return}_{i,t-10}$	1.51***	0.09	2.50***
	(3.47)	(0.12)	(3.73)
$\text{Return}_{i,t-11}$	1.81***	0.06	2.59***
	(3.72)	(0.09)	(3.05)
Observations	18,919	8,799	7,260
R-squared	0.187	0.185	0.181

Table IA4. LLM Extrapolation of "Changes in Cash Flow" Performance

This table presents the results from linear regressions at the contest-stock-week level, as specified in Eq. (1) in the main text. Forcerank is the average ranking of a stock across all participants in a contest, with ten indicating the highest rank and one the lowest. The dependent variable is the performance ranking generated by GPT-4o. The prompts rely on the same lagged weekly stock return data as in the main analysis, but the prompts are revised to describe the data as quarterly changes in scaled cash flows, and the LLM is prompted to predict performance over the next quarter. The explanatory variables include lagged returns (described as changes in cash flows) relative to forecast week *t*+1. Standard errors are clustered by contest, and *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

	$\text{GPT}-\text{4o Rank}_{i,t+1}^{\text{Cash Flows}}$		
	(1)	(2)	(3)
$\text{Forcerank}_{i,t}$		0.28***	0.17***
		(27.11)	(20.51)
$\text{Return}_{i,t}$	39.49***		37.32***
	(37.47)		(36.41)
$\text{Return}_{i,t-1}$	3.86***		3.46***
	(6.09)		(5.67)
$\text{Return}_{i,t-2}$	3.15***		2.77***
	(5.09)		(4.68)
$\text{Return}_{i,t-3}$	1.87***		1.42**
	(3.20)		(2.52)
$\text{Return}_{i,t-4}$	1.44**		1.05*
	(2.29)		(1.72)
$\text{Return}_{i,t-5}$	2.46***		2.03***
	(4.22)		(3.64)
$\text{Return}_{i,t-6}$	0.26		-0.03
	(0.43)		(-0.05)
$\text{Return}_{i,t-7}$	0.78		0.56
	(1.21)		(0.90)
$\text{Return}_{i,t-8}$	1.54**		1.31**
	(2.52)		(2.25)
$\text{Return}_{i,t-9}$	0.81		0.64
	(1.39)		(1.16)
$\text{Return}_{i,t-10}$	2.10***		1.84***
	(3.50)		(3.20)
$\text{Return}_{i,t-11}$	1.72***		1.65***
	(2.86)		(2.86)
$\text{Return}_{i,t-12}$	12,151	12,178	12,151
	0.355	0.080	0.384
Observations	12,606	12,718	12,606
R-squared	0.031	0.004	0.032

Table IA5. LLM Extrapolation Past Returns: Adding Firm Characteristics
This table presents the results from linear regressions at the contest-stock-week level, as specified in Eq. (1) and Table 2. In this implementation, we include a variety of firm characteristics in the forecast prompts. In Specification (1), the prompt adds to lagged returns several market information measures: Shares Outstanding, Market Capitalization, Open Price, Close Price, Day High, Day Low, 52-week high, 52-week low, 10-Day Average Volume, and Beta. Specification (2) considers prompts that include lagged returns and fundamental information measures: Revenue, EBITDA, Earnings Per Share, Gross Margin, Net Margin, ROE, Debt-to-Equity, P/E Ratio, and Book-to-Market. Specification (3) considers prompts that include lagged returns and all firm information measures. Standard errors are clustered by contest, and *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample period is from February 2016 to December 2017.

	$\text{GPT}-4\text{o Rank}^{12w}_{i,t+1}$		
	Market Information Controls	Fundamental Information Controls	Market and Fundamental Information Controls
	(1)	(2)	(3)
$\text{Return}_{i,t}$	37.45***	36.23***	36.73***
	(38.21)	(37.78)	(38.01)
$\text{Return}_{i,t-1}$	6.33***	6.81***	6.51***
	(10.08)	(11.20)	(10.67)
$\text{Return}_{i,t-2}$	4.02***	4.65***	4.14***
	(6.81)	(7.84)	(7.22)
$\text{Return}_{i,t-3}$	3.78***	4.49***	4.19***
	(6.54)	(7.79)	(7.35)
$\text{Return}_{i,t-4}$	3.54***	4.01***	3.72***
	(5.65)	(6.37)	(5.98)
$\text{Return}_{i,t-5}$	4.66***	5.01***	4.63***
	(8.15)	(8.71)	(8.06)
$\text{Return}_{i,t-6}$	2.33***	2.54***	2.15***
	(3.97)	(4.17)	(3.65)
$\text{Return}_{i,t-7}$	3.11***	3.03***	2.74***
	(4.89)	(4.71)	(4.23)
$\text{Return}_{i,t-8}$	2.90***	3.19***	2.75***
	(4.72)	(5.29)	(4.50)
$\text{Return}_{i,t-9}$	2.54***	2.62***	2.79***
	(4.25)	(4.46)	(4.71)
$\text{Return}_{i,t-10}$	3.69***	4.11***	3.66***
	(6.03)	(6.83)	(6.09)
$\text{Return}_{i,t-11}$	4.04***	4.87***	4.52***
	(6.79)	(8.38)	(7.64)
Observations	12,142	12,142	12,142
R-squared	0.343	0.331	0.333

Table IA6. LLM Extrapolation of Price Charts: 24-Week Price Charts

This table presents the results from linear regressions at the contest-stock-week level, as specified in Eq. (1) in the main text. Forcerank is the average ranking of a stock across all participants in a contest, with ten indicating the highest rank and one the lowest. The dependent variable is the performance ranking generated by GPT-4o based on price charts over the past 24 weeks. The explanatory variables include lagged returns relative to forecast week *t*+1. Standard errors are clustered by contest, and *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample period is from February 2016 to December 2017.

	$\text{GPT}-4\text{o Rank}_{i,t+1}^{24w\ Chart}$		
	(1)	(2)	(3)
$\text{Forcerank}_{i,t}$		0.06***	0.03***
		(6.40)	(3.44)
$\text{Return}_{i,t}$	6.36***		5.94***
	(11.67)		(10.81)
$\text{Return}_{i,t-1}$	4.48***		4.40***
	(8.43)		(8.28)
$\text{Return}_{i,t-2}$	3.76***		3.69***
	(7.43)		(7.31)
$\text{Return}_{i,t-3}$	3.75***		3.67***
	(6.74)		(6.63)
$\text{Return}_{i,t-4}$	2.92***		2.84***
	(5.70)		(5.58)
$\text{Return}_{i,t-5}$	3.36***		3.28***
	(6.54)		(6.40)
$\text{Return}_{i,t-6}$	2.51***		2.46***
	(4.79)		(4.71)
$\text{Return}_{i,t-7}$	2.23***		2.18***
	(4.38)		(4.30)
$\text{Return}_{i,t-8}$	2.36***		2.32***
	(4.64)		(4.56)
$\text{Return}_{i,t-9}$	1.85***		1.82***
	(3.66)		(3.61)
$\text{Return}_{i,t-10}$	2.48***		2.42***
	(5.29)		(5.18)
$\text{Return}_{i,t-11}$	2.10***		2.10***
	(4.34)		(4.35)
$\text{Return}_{i,t-12}$	1.38***		1.36***
	(3.11)		(3.07)
$\text{Return}_{i,t-13}$	1.38***		1.34***
	(2.81)		(2.75)
$\text{Return}_{i,t-14}$	1.92***		1.89***
	(3.81)		(3.77)
$\text{Return}_{i,t-15}$	0.95**		0.92**

	(2.05)		(1.99)
$Return_{i,t-16}$	1.05**		1.02**
	(2.21)		(2.15)
$Return_{i,t-17}$	0.46		0.47
	(0.98)		(1.01)
$Return_{i,t-18}$	0.49		0.48
	(1.03)		(1.00)
$Return_{i,t-19}$	1.05**		0.99**
	(2.19)		(2.07)
$Return_{i,t-20}$	0.47		0.43
	(0.97)		(0.89)
$Return_{i,t-21}$	0.49		0.49
	(0.95)		(0.94)
$Return_{i,t-23}$	0.18		0.16
	(0.37)		(0.32)
$Return_{i,t-23}$	0.69		0.69
	(1.44)		(1.44)
Observations	12,606	12,718	12,606
R-squared	0.031	0.004	0.032

Table IA7. Revised Sentiment Prompt based on Greenwood and Shleifer (2014)

Prompt:

Prompt for Revising an Existing Prompt Using an Academic Study

You are an expert in prompt engineering and behavioral decision-making. Your task is to revise an original prompt so that it produces more accurate or informed responses, based on the insights of an academic study.

Instructions:

1. **Read the Original Prompt.** Understand the task it is asking a language model to perform.
2. **Read the Academic Study.** Extract key behavioral, informational, or structural findings that can improve how a language model reasons or makes predictions in the context of the original prompt.
3. **Generate a Revised Prompt** that:
    - Retains the structure and intention of the original prompt.
    - Integrates relevant lessons or adjustments derived from the academic study.
    - If helpful, adds background information, rephrases questions, or inserts reasoning steps to reduce errors, bias, or misinterpretation.
    - Optionally includes a short rationale (in comments) explaining the changes made.

Input:

Original Prompt:

[The csv data contain the monthly stock returns in months t-12 to t-1.
Please answer the following questions:
Do you feel the direction of the stock market over the next six months will be up (bullish), no change (neutral) or down (bearish)?
How confident are you about this prediction?
Your output will be in json format with the following format:
'{"prediction":,"confidence":}'. 1 stands for bullish, 0 for neutral, and -1 for bearish. Confidence represents a probability that ranges from 0 to 1.]

Academic Study: [*Greenwood and Shleifer (2014), "Expectations of Returns and Expected Returns"*]

Output:

Revised Prompt: [Updated prompt, ready to use]

Table IA8. Market Return Extrapolation: Alternative LLMs

The table repeats the analysis in Specification (2) of Table 8 using alternative large language models to generate market sentiment based on US stock market (S&P 500) returns over the past 12 months. Specification (1) reports the baseline GPT-4o results from Table 8. In Specification (2), sentiment is generated by GPT-o1 preview; in Specification (3), sentiment is generated by Claude 3.5 Sonnet, and in Specification (4), we use Gemini 2.5 Pro to generate sentiment measures.

	GPT-4o	GPT-o1	Claude 3.5	Gemini 2.5
	(1)	(2)	(3)	(4)
S&P Return_{t}	9.17***	8.94***	10.64***	12.04***
	(16.68)	(13.30)	(17.74)	(17.16)
S&P Return_{t-1}	5.38***	7.65***	5.74***	6.26***
	(10.20)	(10.78)	(9.83)	(8.97)
S&P Return_{t-2}	3.99***	4.64***	3.60***	3.82***
	(8.78)	(7.46)	(6.52)	(6.34)
S&P Return_{t-3}	3.65***	2.27***	2.61***	1.70***
	(7.48)	(3.59)	(4.85)	(3.02)
S&P Return_{t-4}	2.68***	2.49***	1.92***	0.88
	(4.89)	(3.99)	(3.34)	(1.37)
S&P Return_{t-5}	3.39***	2.32***	2.31***	0.93
	(7.04)	(3.92)	(3.95)	(1.43)
S&P Return_{t-6}	2.48***	1.65***	1.29**	1.21**
	(6.03)	(2.61)	(2.30)	(2.07)
S&P Return_{t-7}	2.00***	-0.44	1.43***	0.48
	(4.43)	(-0.76)	(2.87)	(0.91)
S&P Return_{t-8}	1.78***	1.42**	1.43***	0.98*
	(3.69)	(2.27)	(2.63)	(1.71)
S&P Return_{t-9}	2.13***	0.78	1.05**	1.26**
	(4.83)	(1.22)	(2.00)	(2.07)
S&P Return_{t-10}	1.47***	0.31	1.17**	0.72
	(3.26)	(0.46)	(2.49)	(1.23)
S&P Return_{t-11}	2.12***	0.41	0.66	0.17
	(5.11)	(0.73)	(1.37)	(0.32)
Observations	438	438	438	438
R-squared	0.687	0.534	0.630	0.593

Table IA9. Market Return Extrapolation: Simulated Returns with Varying Persistence

This table presents the results from linear regressions of sentiment measures on lagged aggregate returns. The dependent variable is the GPT-4o sentiment generated based on the simulated aggregate market returns for the past 12 months. For each month between 1927 and 2004, we construct simulated market returns using an AR(1) process that matches the mean and standard deviation of the market's monthly returns over the preceding 12 months. We consider seven different autocorrelation values that correspond to the different specifications: -0.9, -0.6, -0.3, 0, 0.3, 0.6, and 0.9. We then repeat the GPT-4o market sentiment prompts using the simulated return data and repeat the analysis in Table 8. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

	$\text{GPT}-4\text{o Sentiment}_{t+1}$						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Autocorrelation	-0.9	-0.6	-0.3	0.0	0.3	0.6	0.9
$\text{S\&P Return}_{t}$	5.87***	6.26***	6.83***	6.43***	6.70***	6.54***	6.58***
	(9.49)	(16.82)	(19.33)	(15.72)	(15.04)	(15.03)	(16.52)
$\text{S\&P Return}_{t-1}$	4.44***	4.50***	4.38***	4.01***	4.15***	4.73***	3.08***
	(5.93)	(9.68)	(10.42)	(9.55)	(11.06)	(11.73)	(5.50)
$\text{S\&P Return}_{t-2}$	4.19***	3.13***	3.49***	4.15***	3.46***	2.67***	3.85***
	(6.66)	(7.58)	(10.89)	(11.83)	(7.76)	(6.64)	(7.45)
$\text{S\&P Return}_{t-3}$	3.75***	2.75***	3.58***	3.45***	3.19***	4.05***	2.06***
	(4.84)	(6.72)	(7.98)	(8.71)	(7.44)	(9.74)	(4.75)
$\text{S\&P Return}_{t-4}$	3.34***	2.61***	2.84***	2.86***	2.38***	2.05***	1.65***
	(5.28)	(6.12)	(8.28)	(8.99)	(5.96)	(4.61)	(3.27)
$\text{S\&P Return}_{t-5}$	2.43***	2.35***	2.16***	2.27***	2.54***	1.88***	1.68***
	(4.33)	(6.27)	(5.17)	(8.17)	(6.18)	(4.65)	(3.17)
$\text{S\&P Return}_{t-6}$	1.62***	2.42***	2.13***	1.62***	1.87***	1.94***	0.87*
	(2.88)	(5.53)	(4.74)	(4.37)	(4.74)	(4.31)	(1.78)
$\text{S\&P Return}_{t-7}$	1.91***	2.12***	1.71***	2.26***	1.66***	0.42	0.79*
	(3.78)	(4.41)	(5.92)	(6.85)	(4.17)	(0.85)	(1.68)
$\text{S\&P Return}_{t-8}$	2.08***	2.10***	2.57***	1.22**	1.03***	1.88***	0.29
	(3.56)	(5.27)	(8.09)	(2.56)	(2.78)	(4.28)	(0.56)
$\text{S\&P Return}_{t-9}$	1.22**	1.79***	2.55***	0.72***	1.60***	1.02***	0.76
	(2.33)	(5.29)	(9.09)	(2.75)	(4.17)	(2.63)	(1.43)
$\text{S\&P Return}_{t-10}$	0.88	1.61***	0.91**	2.18***	0.70*	0.77*	0.56
	(1.64)	(3.77)	(2.58)	(5.60)	(1.83)	(1.68)	(1.11)
$\text{S\&P Return}_{t-11}$	1.78***	1.25***	2.02***	1.72***	1.37***	1.41***	0.39
	(3.73)	(3.35)	(5.70)	(4.56)	(3.84)	(3.59)	(0.91)
Observations	1,164	1,164	1,164	1,164	1,164	1,164	1,164
R-squared	0.629	0.635	0.661	0.617	0.610	0.599	0.598

Table IA10. Revised Confidence Interval Prompt based on Hartzmark and Sussman (2024)

Prompt:

Prompt for Revising an Existing Prompt Using an Academic Study
You are an expert in prompt engineering and behavioral decision-making. Your task is to revise an original prompt so that it produces more accurate or informed responses, based on the insights of an academic study.

Instructions:

4. **Read the Original Prompt.** Understand the task it is asking a language model to perform.
5. **Read the Academic Study.** Extract key behavioral, informational, or structural findings that can improve how a language model reasons or makes predictions in the context of the original prompt.
6. **Generate a Revised Prompt** that:
    - Retains the structure and intention of the original prompt.
    - Integrates relevant lessons or adjustments derived from the academic study.
    - If helpful, adds background information, rephrases questions, or inserts reasoning steps to reduce errors, bias, or misinterpretation.
    - Optionally includes a short rationale (in comments) explaining the changes made.

Input:

Original Prompt:
[Below are the monthly returns for a financial asset over the past 120 months.
Please answer the following questions on next month's return. There is a 1-in-10 chance the actual return will be less than a%. I expect the next month's return to be: b%. There is a 1-in-10 chance the actual return will be greater than c%. Please return a JSON object in the following format:
'{"low": a%,"expected": b%,"high": c%}'.]
Academic Study: [*Hartzmark and Sussman (2024), "Eliciting Expectations"*]

Output:

Revised Prompt: [Updated prompt, ready to use]

Table IA11. Prompt for Generating Return Bin Forecasts

Prompt:

Below are the monthly returns for a financial asset over the past 120 months. Please consider the upcoming monthly return for this asset. Indicate how likely you believe it is that the return will fall into each of the following categories. Your responses should be whole numbers that sum to 100, representing a full distribution of probabilities.

Bin	Return Range (%)
1	**r < –30**
2	**–30 ≤ r < –18**
3	**–18 ≤ r < –10**
4	**–10 ≤ r < –3**
5	**–3 ≤ r < 3**
6	**3 ≤ r < 10**
7	**10 ≤ r < 18**
8	**18 ≤ r < 30**
9	**r ≥ 30**

Return only a JSON object in this exact schema:

```
{ "bin_1": ,
  "bin_2": ,
  "bin_3": ,
  "bin_4": ,
  "bin_5": ,
  "bin_6": ,
  "bin_7": ,
  "bin_8": ,
  "bin_9": }
```