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*

January 2025

Abstract

We examine how large language models (LLMs) interpret historical stock returns and price charts when prompted to forecast stock returns over horizons from one week to one month. While stock returns exhibit short-term reversals, LLM forecasts over-extrapolate, placing excessive weight on recent performance similar to humans. LLM forecasts appear optimistic relative to historical and future realized returns. When prompted for 80% confidence interval predictions, LLM responses are better calibrated than survey evidence but pessimistic about outliers, leading to skewed forecast distributions. The findings suggest LLMs manifest common behavioral biases when forecasting expected returns but are better at gauging risks than humans.

JEL: D84, G17, G40, O33

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

*Chen and Gulen are from the Mitch Daniels School of Business, Purdue University, chen4144@purdue.edu; hgulen@purdue.edu. Green is from Goizueta Business School, Emory University, clifton.green@emory.edu. Zhou is from Zicklin School of Business, Baruch College, dexin.zhou@baruch.cuny.edu. We thank seminar participants at Syracuse University, Clemson University, McMaster University, and Renmin University, for 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), firm auditors (Fedyk et al., 2022), and debt collectors (Choi et al., 2024). On the other hand, large language models (LLMs) 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).[1]

To the extent that AI algorithms mimic human decision-making, they may also incorporate cognitive biases that individuals exhibit in financial contexts. An extensive literature documents that people excessively extrapolate from recent performance, tend to be overly optimistic, and are unreasonably confident in their predictions.[2] In this study, we assess the extent to which state-of-the-art generative AI models, as proxied by OpenAI's large language model ChatGPT-4o, manifest behavioral biases when prompted to provide stock return forecasts based on historical return data and stock price chart images.

While "Large Language Model" might suggest a primary focus on textual data, these models possess capabilities that extend well beyond language processing. For instance, ChatGPT can interpret and analyze visual elements such as plots and charts, demonstrating skill in handling visual and numeric data. We begin by examining how LLMs interpret the timing of historical

[1] 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).
[2] Hirshleifer (2015) and Barberis (2018) provide reviews.

return data. Utilizing survey data, 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., 2023).

In one study, 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 for a set group of stocks. In our first AI investigation, we query ChatGPT to "compete" in each of the 1,379 stock ranking contests while providing 12 weeks of historical return data for the ten stocks in each contest.[3]

Consistent with DHJ, we confirm that human performance rank forecasts place positive weights on all 12 historical returns, with the greatest positive emphasis on the previous week and each preceding week becoming generally less important. In stark contrast to human predictions, realized return regressions for stocks used in these contests show evidence of weekly return reversals, with negative coefficients on the past four return lags, including one significant coefficient. This suggests that Forcerank participants' forecast bias likely emanates from over-extrapolative expectation formation.

It is unclear what relation will emerge between ChatGPT forecasted returns and historical return data. While humans are prone to over-extrapolation, the phenomenon of short-term return reversals has been well-known for decades. For example, Jegadeesh (1990) and Lehman (1990) document significant profits using short-term reversal strategies, and together they have received

[3] Returns are presented without firm identifiers to prevent bias from the training sample, and for robustness, we also analyze simulated contests using returns from after the launch of the model.

more than 6,000 Google Scholar citations. It is possible that ChatGPT's training may incorporate stylized facts about short-term stock reversals. In addition, our query provides exclusively numeric data rather than human-authored text, which may place the algorithm in a less behavioral, more mathematical context. Although LLMs are not specifically designed to handle numeric tasks, their ability to recognize patterns from visual or numerical data, learn from statistical correlations, and understand context allows them to effectively approximate numeric operations.[4]

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 0.279, which suggests a significant commonality between the two forecasts. The regression evidence reveals that both humans and GPT-4o forecasts rely on past data in surprisingly similar ways. As with human forecasts, the coefficients on lagged returns are positive, with the first lag being the largest, the second lag the next largest, the third lag the next largest, and the remaining coefficients noticeably smaller. In light of the empirically observed short-term return reversals, LLM's extrapolation is counterproductive and produces rankings that are negatively associated with future performance.

Participants in the Forcerank contests extrapolate negative and positive returns asymmetrically, consistent with neuroscience studies that show gains and losses are processed by different regions of the brain (Kuhnen and Knutson, 2005). Specifically, humans tend to place greater emphasis on negative returns, and negative performance has a longer-lasting effect on expectations (Gulen and Lim, 2024). In contrast to the evidence for human subjects, we find that GPT-4o forecasts plae greater emphasis on recent positive returns than negative returns while extrapolating more distant negative returns in ways similar to humans.

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

We next shed light on 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 ChatGPT is prompted to issue performance rank forecasts for the following week. In line with the prompts that include numerical return data, ChatGPT's forecasts continue to extrapolate from past returns when visually inferring return information from the price charts.

The Forcerank setting emphasizes cross-sectional variation in return performance. We also explore how ChatGPT forms aggregate market return forecasts using rolling windows of monthly historical return data for the S&P 500 Index. We compare LLM's aggregate return 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 inferred from survey data, we find that GPT-4o's return forecasts place the largest positive weights on recent returns.

We also consider forecasts from alternative LLMs. In particular, we repeat the S&P 500 return-based queries using the Claude large language model from Anthropic. Although Claude and ChatGPT were trained independently, the correlation between the two return measures is 0.78, suggesting that the same underlying expectation formation behavior is manifested in their responses. The extrapolative coefficients from the Claude forecasts are very similar to the estimates from ChatGPT, suggesting that return extrapolation in LLM forecasts is not confined to individual stocks or a single LLM.[5]

[5] We also find similar patterns with ChatGPT-o1 (released May 2024).

In our next set of queries, we investigate how LLMs react to the relative magnitudes of individual historical returns when predicting distinct characteristics of the return distribution. Ben-David, Graham, and Harvey (2013) (BGH) survey CFOs regarding their projections of the overall market and their individual companies and find evidence of significant miscalibration, with realized returns falling outside the CFO's predicted 80% confidence intervals substantially more than the expected 20%. Hartzmark and Sussman (2024) also find evidence of poorly calibrated forecasts, with US and Japanese investors providing market return forecast intervals that fail to meaningfully reflect the true range of outcomes.

In order to evaluate the calibration of LLM stock return forecasts, we randomly select 10,000 stock-month observations from the 1926-2023 sample period. For this set, we gather ten years of historical monthly return observations and prompt ChatGPT to answer questions similar to the survey in BGH. We first examine whether LLM expected return forecasts appear biased relative to realized outcomes. Humans tend to be overly optimistic in a variety of settings (e.g., Van den Steen, 2004; Hartzmark and Sussman, 2024). If overoptimism manifests in the training data, then LLM forecasts may also be higher than realized returns.

Empirically, we find consistent evidence that LLM expected returns are significantly higher than both historical means and realized returns. For example, the average monthly cross-sectional historical return provided to ChatGPT is 1.4%, and the cross-sectional average for the next period's realized return is 1.15%. Yet the average expected return projected by GPT-4o is roughly twice as large at 2.2%. The observed ChatGPT positive bias is partially attributable to return forecasts being largely truncated at zero, suggesting that the training data may have embedded the economic idea that expected returns should be nonnegative.

Turning to low and high forecasts, we observe that next month's realized return value lies within the GPT-4o 80% confidence interval 76.9% of the time, which is less accurate than the 79.0% that could be contained by simply using the 10^{th} and 90^{th} historical percentiles as the forecasts. However, the evidence of miscalibration in ChatGPT forecasts is much less severe than in the executive surveys, consistent with enhanced numeracy for LLMs. Nevertheless, we observe that the 10^{th} percentile LLM forecast (Low) is significantly less than the 10^{th} historical percentile, suggesting pessimism in projecting unfavorable outcomes. On the other hand, the 90^{th} percentile forecast (High) is also significantly less than the 90^{th} historical percentile.

To deepen our understanding of how LLMs translate historical returns into forecasts, we regress LLM return forecasts on historical 10^{th}, 20^{th}, …, 90^{th} percentiles. Unsurprisingly, when the dependent variable is the forecast of next period's return, the regression produces significant loadings on each of the nine percentile measures. However, we observe that the largest loading is on the 90^{th} percentile, consistent with forecast overoptimism. Examining how Low and High forecasts incorporate historical data, we observe that both forecasts load significantly on the corresponding percentiles, but they also load significantly on percentiles on the other side of the distribution with negative signs, indicating underlying assumptions about distributional symmetry. However, the High forecast is less sensitive to high percentiles than the Low forecast is to low percentiles, suggesting underlying pessimism about the tails of the distribution.

Taken together, the analysis demonstrates that LLM stock return forecasts exhibit over-extrapolation of historical return performance. While LLM forecasts are considerably better calibrated than human forecasts, indicating improved assessments of risk, LLM forecasts tend to be excessively optimistic when predicting expected performance and slightly pessimistic about the

tails of the distribution. As a result, LLM forecasted return distributions are positively skewed compared to historical data.

The findings contribute to recent literature that examines the extent to which LLMs reproduce human behavior in financial contexts.[6] 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 Fedyk et al. (2024) and Lim (2024) find evidence that ChatGPT embeds investment preferences that vary across gender, income, and age. Chen et al. (2023) finds evidence that LLMs are more rational than humans when choosing risky assets. Evidence on using LLMs to generate return forecasts is mixed. Lopez-Lira and Tang (2023) finds that ChatGPT can successfully forecast daily stock returns using news headlines, and Kim, Muhn, and Nikolaev (2024) and Jha et al. (2024) find ChatGPT excels at distilling corporate disclosures, suggesting 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 interpret even straightforward numeric data with fully rational statistical rigor. This highlights the importance of addressing potential biases as AI becomes increasingly integrated into financial decision-making. The current generation of LLMs straddles the line between mimicking human behavior and achieving rational, statistical precision,

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

occupying an intermediate space that reflects both the promise and limitations of these models in their current form.

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 ChatGPT-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, 2022 for more details).[7]

The sample contains 1,283 weekly contests. including 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

[7] 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.

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 LLM forecasts to similar ranks submitted by humans. DHJ examine extrapolative behavior by analyzing how average Forcerank scores load on twelve weeks of lagged stock returns. In our main analysis, we similarly 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:[8]

> 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.[9] Confidence represents a probability that ranges from 0 to 1.[10]

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). 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. Additionally, for robustness, we also analyze data after the December 2023 cutoff for GPT-4o (see Section 4.2).

[8] 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/

[9] 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.

[10] ChatGPT'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.

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 ChatGPT market sentiment measure is -1 for bearish, 0, for neutral, and 1 for bullish.

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

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

> 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.

2.3 CEO Calibration Surveys

Prior analyses largely focus on the direction or ranking of future stock returns. While these prompts are helpful in understanding the expectation of future returns, they are not informative about LLMs' predictions of the future return distributions. Thus, in addition to examining expected returns, we also seek to understand how LLMs determine other aspects of the forecasted return distribution. We are guided by the CFO survey examined in Ben-David, Graham, and Harvey (2013). The Duke CFO has been surveying executives since 1996 and along with overall outlook questions, the survey includes additional questions that can vary over time (e.g., questions related to upcoming presidential elections). BGH examines survey vintages that ask executives to forecast actual aggregate market returns as well to provide the 80% forecast confidence intervals. Motivated by BGH, we prompt ChatGPT to issue 10^{th} and 90^{th} percentile forecasts in addition to expected returns.

We deviate from the executive survey by asking ChatGPT to produce stock-level return forecasts instead of aggregate market forecasts. Stock-level data offers greater variation in the historical return distributions, and it also permits many more observations than a single aggregate market return series can provide. In our approach, we randomly select 100 months from the 1926

to 2023 time period. For each selected month, we choose 100 stocks, with 10 from each size decile. For this set of 10,000 stock-months, we gather up to ten years of historical monthly returns (requiring no fewer than five years of returns). We place each set of observations in a .csv file and prompt GPT-4o with the following text guided by the survey questions in BGH.

> 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 1,286 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 Investor 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 ChatGPT 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

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 (essentially smaller neural networks) to analyze when small groups of neurons fire together, creating mind maps 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 return 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 low (10th percentile) and high (90th percentile) forecasts represent distinct LLM features that potentially weigh returns differently.

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 ChatGPT-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 ChatGPT-4 forecasted rank for the same contest-stock based on returns *t* through *t*-1; 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.

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 ChatGPT 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 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-squares are considerably lower for humans than for ChatGPT (3.4% vs 35.2%) 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 ChatGPT 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,[11] and in direct contrast to human and GPT-4o expectations, realized weekly returns

[11] 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).

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.[12]

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 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 ChatGPT 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

[12] 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.

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 ChatGPT 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 ChatGPT-4 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 (2) that the strongest extrapolation in ChatGPT 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 ChatGPT 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 0.82.

4.1.3. LLM Extrapolation from Price Charts

Large language models can interpret images as well as numerical data, and we next examine how ChatGPT 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 ChatGPT-4 while prompting for next-week performance ranks. We then repeat the extrapolation regression using price chart-based performance rank forecasts.

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 5 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 IA4. The price chart evidence suggests that the inclination for LLMs to extrapolate from recent returns extends beyond numeric data to unstructured image data.

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.

We run three additional prompts for each contest where we include, alongside the l2 weeks of lagged stock returns for each stock, and 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 37.5 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.

4.1.4. 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[13]:

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

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

where $w_s = \frac{\lambda_2^s}{\sum_{j=0}^{n} \lambda_2^j}$ and $Y_{i,t+1}$ 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 $\lambda_1(1 - \lambda_2)$.

Table 6 presents the regression estimates for the exponential decay model. The level coefficient (λ_1) is larger for GPT-4o forecast ranks than for humans, but this is perhaps to be expected since humans have other sources of information at their disposal to shape their expectations. Moreover, the slope coefficient (λ_2) is smaller for GPT-4o, suggesting an even greater weight on recent past returns, resulting in a higher degree of extrapolation than human forecasts (38.03 vs 12.19 for humans). Results from ChatGPT ranks generated based on 24 lagged weekly returns are similar.

4.1.5. 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, 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 Eq. (1). In other words, *Predicted GPT-4o* is the weighted average of the past 12 weekly returns that best predicts the LLM forecast, and the residual of this regression 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 7 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 Extrapolation in Aggregate Market Sentiment

The Forcerank setting analyzes the relative performance in a cross section of stocks. In this section, we examine GPT-4o expectations measures of aggregate market performance to survey evidence from the American Association of Individual Investor survey.

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

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.[14] For ChatGPT, $\text{Sentiment}_{i,t+1}$ is equal to 1 if the GPT-4o 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, we find that ChatGPT 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 weeks. In Specification (5), we examine the extent to which future six-month returns are related to lagged monthly returns. None of the coefficients are significant, suggesting little support for LLMs' extrapolative approach to forecasting based on past returns.

4.2.1 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

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

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) obtained from the prompt "Summarize the [included] academic study in 500 words. The text will be used to help guide LLM decision-making."[15]

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, 0.761, 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 5.55. 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, over extrapolation 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.

4.2.2 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 first repeat the market sentiment queries using another LLM that is widely regarded at the time of the analysis, the Claude 3.5 Sonnet model from Anthropic. Table IA4 in the Internet

[15] The prompt and resulting summary is reported in Table IA6 in the Internet Appendix.

Appendix presents the results. We observe that the two different LLM sentiment forecasts are closely related. Regressing the Claude sentiment measure on the ChatGPT sentiment produces a coefficient of 0.81 and an R-squared of 0.603, suggesting a correlation of 0.78 between the two sentiment measures. Moreover, the magnitudes of the lagged return coefficients are very similar when comparing Specification (2) of Table 8 to Specification (1) of Table IA7.

We also consider the most recent model by OpenAI, released in May of 2024, which is labeled o1. In Table IA8 in the Internet Appendix, we repeat the analysis using the o1 model and again find similar results, with the correlation between GPT-4o Sentiment and o1 Sentiment being 0.71. Moreover, both sentiment measures produce similar weights on lagged market returns in the extrapolative regressions.

4.2.3 Sensitivity of LLM Extrapolation to Underlying Autocorrelation

Afrouzi et al. (2003) 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 sentiment analysis evidence helps confirm that return extrapolation in ChatGPT forecasts is not confined to individual stocks or a specific large language model, easy to eliminate through simple prompt engineering, or sensitive to the level of underlying autocorrelation in the data.

5. Bias and Miscalibration in Large Language Model Forecasts

In this section, we ask GPT-4o to provide specific forecasts about the characteristics of the stock return distribution. Our guiding setting is the company executive survey analyzed in Ben-David, Graham, and Harvey (2013), which asks CFOs to forecast actual returns as well as provide 80% forecast confidence intervals. Motivated by BGH, we provide monthly historical return data and prompt GPT-4o to issue next-period return forecasts in addition to the 10^{th} and 90^{th} percentiles for a random sample of 10,000 stock-months.

5.1 Univariate Evidence

We first examine 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 more likely to choose stocks for which they have overestimated the likelihood of success, similar to the winner's curse. If overoptimism manifests in the training data, then LLM forecasts may also be higher than realized returns. Table 10 presents forecast statistics for the sample of 9,959 stock-months that survived the historical return requirements. We observe that the cross-sectional mean of the GPT-4o forecast for next month's stock return is 2.2%. The LLM forecast is considerably larger than the average historical mean for the data that were provided in the prompt (1.4%) and roughly twice the magnitude of next month's realized return (1.1%), with both differences being statistically significant.

Figure 2 (middle 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 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 (0.45% of forecasts), which suggests that the LLM's training may have embedded the idea that expected returns should be nonnegative. However, Table 11 indicates that the median, 75^{th}, and 90^{th} percentiles of the distribution of GPT-4o expected return forecasts are all higher than the average historical equivalents, suggesting that ChatGPT's positive expected return bias extends beyond truncating at zero.

We next examine LLM forecasts of Low and High returns. The average GPT-4o 80% confidence interval is 23.4%, which is smaller than the average historical confidence interval of 25.5%. We observe that next month's realized return value lies within the GPT-4o confidence

interval 76.9% of the time, which is less accurate than the 79.0% that could be contained by simply using the 10th and 90th historical percentiles as the forecasts. The miscalibration is primarily on the upside, with 12.7% of realized returns occurring above the High forecast, compared to 10.4% on the downside.

The evidence of miscalibration in ChatGPT forecasts is much less severe than in the executive surveys, with BGH reporting that realized market returns are within executives' 80% confidence intervals only 36% of the time. The findings of substantially better calibration of GPT-4o forecasts relative to humans are consistent with improved numeracy. Despite the evidence of generally being well-calibrated, the GPT-4o Low forecast is significantly less than the 10th historical percentile, suggesting conservatism in projecting unfavorable outcomes. On the other hand, the GPT-4o High forecast is also significantly less than the 90th historical percentile. These patterns are evident in Figure 3, which plots the distribution of Low – historical 10th percentile, Forecast – historical mean, and High – historical 90th percentile.

5.2 Regression Evidence

To deepen our understanding of how LLMs translate historical returns into forecasts, we regress return ChatGPT-4 forecasts on characteristics of the historical return distribution as follows:

$$\text{Forecast}_{i,t+1} = \gamma_0 + \beta_1 Ret_i^{min} + \beta_2 Ret_i^{10\%} + \cdots + \beta_{10} Ret_i^{90\%} + \beta_{11} Ret_i^{max} + \varepsilon_{i,t}. \quad (5)$$

$\text{Forecast}_{i,t+1}$ is either the GPT-4o forecast of the next period return, the Low (10th percentile) forecast, or the High (90th percentile) forecast. Ret_i^{min} and Ret_i^{max} are the minimum and maximum of the historical return distribution provided in the prompt, and $Ret_i^{10\%}$ is the 10th percentile of the historical return data, etc. The results are presented in Table 11.

Intuitively, when the dependent variable is the expected return forecast, the regression returns significant loadings on each of the nine percentile measures. However, we observe that the largest loading is on the 90th percentile, providing additional evidence of positive bias in return forecasts. Examining how *Low* and *High* forecasts incorporate historical data, we observe that both forecasts load significantly on the corresponding percentiles, but they also load significantly on percentiles on the other side of the distribution with negative signs, indicating underlying assumptions about distributional symmetry. However, the *High* forecast is less sensitive to high percentiles than the *Low* forecast is to low percentiles, suggesting underlying pessimism about the tails of the distribution. The evidence that Low and High forecasts have different characteristics suggests that the training data has led ChatGPT to treat high and low stock return outcomes as distinct model features.

Taken together, the findings suggest that LLM forecasts are considerably better calibrated than human forecasts, indicating improved assessments of risk. However, LLM forecasts tend to be excessively optimistic when predicting expected performance and slightly pessimistic about the tails of the distribution. As a result, LLM forecasted return distributions are positively skewed compared to historical data.

3.3 Discussion

Our analysis suggests that LLM stock performance forecasts exhibit excessive extrapolative behavior, tend to be overoptimistic regarding expected returns, and are downward-biased when forecasting the tails 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," ChatGPT-4's response includes Extrapolation Bias, Recency Bias, Overconfidence, Confirmation Bias, Hindsight Bias, and the Availability Heuristic, and it offers definitions of each. ChatGPT-4 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 outputs 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 by retraining on a more specific datasets tailored to a particular setting, which can improve performance for domain tasks. Moreover, LLMs can also be encouraged to generate code when presented with data, further improving their data analysis capabilities. However, when an LLM faces a task without well-answered examples in the fine-tuning data, general knowledge that is susceptible to biases will likely play a larger role.

On one hand, closely aligning Generative AI model behavior with the nuances of 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). Conversely, when Generative AI models are used in expert capacities, such as serving as financial advisors, neutrality, rationality, and data-driven

insights are paramount. Expert contexts demand a systematic removal of bias to ensure informed and objective outputs. Currently, LLMs embody a blend of these human-like and expert characteristics, reflecting the complexities of their training. Our findings underscore the importance of understanding how these models align with human behavior and highlight the need for targeted development to guide LLMs towards 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 of embedding detrimental cognitive biases that are present in human decision-making. This study examines whether OpenAI's GPT-4 manifests behavioral biases when provided with historical return data and price chart images and prompted for stock return forecasts. In particular, we assess how ChatGPT reacts to the timing and relative magnitude of individual historical returns when forming return forecasts.

Our empirical analysis indicates that ChatGPT and human forecasts rely on past data in similar ways, with a positive, gradually declining 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 excessive extrapolative 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.

Additionally, ChatGPT tends to predict higher returns than both historical means and realized outcomes, indicating that its training data may have embedded an overly positive outlook on stock performance. In contrast, LLMs appear more pessimistic when forecasting the tails of the return distribution. We observe that while forecasts of 10^{th} percentile outcomes were more negative when compared with historical data, 90^{th} percentile forecasts were also lower than the upper percentiles of the historical distribution, suggesting a skewed interpretation of potential risks and rewards.

The findings contribute to the broader discourse on AI integration in financial decision-making, highlighting the need to address potential biases in LLM-generated forecasts. While LLMs show promise in assessing risk, the study cautions against assuming that these models interpret numeric data with fully rational statistical rigor. Our analysis adds to the growing literature on how LLMs replicate human behavior in financial contexts and underscores the importance of critically evaluating AI's role in finance.

Appendix A

A.1 Forecast Variables

- $\text{Forecerank}_{i,t}$ – 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.
- $\text{ChatGPT Rank}_{i,t}$ – ChatGPT-4's ranking of the performance of stock *i* for week *t* when provided with historical return data. Source: ChatGPT-4 prompts.
  - $\text{ChatGPT Rank}_{i,t}^{12w}$ – 12 weeks of historical return data are included in the prompt.
  - $\text{ChatGPT Rank}_{i,t}^{24w}$ – 24 weeks of historical return data are included in the prompt.
  - $\text{ChatGPT Rank}_{i,t}^{12w\ Chart}$ – An image of a price chart with 12 weeks of historical return data is included in the prompt.
  - $\text{ChatGPT Rank}_{i,t}^{24w\ Chart}$ – An image of a price chart with 24 weeks of historical return data is included in the prompt.
- Predicted $\text{ChatGPT Rank}_{i,t}$ – The fitted value obtained from regressing ChatGPT rank on lagged returns as in Specification (2) of Table 5
- Residual $\text{ChatGPT Rank}_{i,t}$ – The residual value obtained from regressing ChatGPT 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.
- $ChatGPT\ Sentiment_t$ – ChatGPT-4'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: ChatGPT-4 prompts.
- $\text{ChatGPT Low}_{i,t}$ – ChatGPT-4'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: ChatGPT-4 prompts.
- $\text{ChatGPT Expected}_{i,t}$ – ChatGPT-4'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: ChatGPT-4 prompts.
- $\text{ChatGPT High}_{i,t}$ – ChatGPT-4'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 and no fewer than five) years of monthly stock returns. Source: ChatGPT-4 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 *t* in week *t*. The rank is from 1 to 10 for the ten stocks in the Forcerank contest.
- 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 sum of the market value of equity and total debts, expressed as a fraction of total assets, 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.

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(3), pp.337-351.

Atmaz, A., Gulen, H., Cassella, S. and Ruan, F., 2023. Contrarians, Extrapolators, and Stock Market Momentum and Reversal. *Management Science*.

Avramov, D., Chordia, T. and Goyal, A., 2006. Liquidity and Autocorrelations in Individual Stock Returns. *The Journal of Finance*, 61(5), pp.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, pp. 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), 1-24.

Bartlett, R., Morse, A., Stanton, R. and Wallace, N., 2022. Consumer-Lending Discrimination in the FinTech Era. *Journal of Financial Economics*, 143(1), pp.30-56.

Ben-David, I., Graham, J.R. and Harvey, C.R., 2013. Managerial Miscalibration. *The Quarterly Journal of Economics*, 128(4), pp.1547-1584.

Ben-Rephael, A., Da, Z. and Israelsen, R.D., 2017. It Depends on Where you Search: Institutional Investor Attention and Underreaction to News. *The Review of Financial Studies*, 30(9), pp.3009-3047.

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

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.

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

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, 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*.

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(2), pp.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(1), pp.175-196.

Da, Z., Liu, Q. and Schaumburg, E., 2014. A Closer Look at the Short-Term Return Reversal. *Management Science*, 60(3), pp.658-674.

Das, S., Kuhnen, C.M. and Nagel, S., 2020. Socioeconomic Status and Macroeconomic Expectations. *The Review of Financial Studies*, 33(1), pp.395-432.

Daniel, K., M. Grinblatt, S. Titman, and R. Wermers, 1997, Measuring Mutual Fund Performance with Characteristic-Based Benchmarks, *Journal of Finance* 52, 1035-1058.

Fedyk, A., Hodson, J., Khimich, N. and Fedyk, T., 2022. Is Artificial Intelligence Improving the Audit Process? *Review of Accounting Studies*, 27(3), pp.938-985.

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(1), pp.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.

Garcia, D., Hu, X. and Rohrer, M., 2023. The Colour of Finance Words. *Journal of Financial Economics*, 147(3), pp.525-549.

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(3), pp.714-746.

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

Gulen, H. and Woeppel, M., 2023. Extrapolation Bias and Short-Horizon Return Predictability. Kelley School of Business Research Paper, (2021-09).

Hartzmark, S.M., Hirshman, S.D. and Imas, A., 2021. Ownership, Learning, and Beliefs. The *Quarterly Journal of Economics*, 136(3), pp.1665-1717.

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

Hirshleifer, D., 2015. Behavioral Finance. In *Annual Review of Financial Economics*, 7(1), pp.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(3), pp.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(6), pp.3193-3249.

Kim, A., Muhn, M. and Nikolaev, V.V., 2024. Bloated Disclosures: Can ChatGPT Help Investors Process Information? Chicago Booth Research Paper, (23-07), pp.2023-59.

Kuchler, T. and Zafar, B., 2019. Personal Experiences and Expectations about Aggregate Outcomes. *The Journal of Finance*, 74(5), pp.2491-2542.

Kuhnen, C.M., 2015. Asymmetric Learning from Financial Information. *The Journal of Finance*, 70(5), pp.2029-2062.

Kuhnen, C.M. and Knutson, B., 2005. The Neural Basis of Financial Risk Taking. *Neuron*, *47*(5), pp.763-770.

Lehmann, B. N, 1990. Fads, Martingales, and Market Efficiency. *The Quarterly Journal of Financial Economics*, 105(1), pp.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(2), pp.254-266.

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

Obaid, K. and Pukthuanthong, K., 2022. A Picture Is Worth a Thousand Words: Measuring Investor Sentiment by Combining Machine Learning and Photos from News. *Journal of Financial Economics*, 144(1), pp.273-297.

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

Van den Steen, E., 2004. Rational Overoptimism (and Other Biases). *American Economic Review*, 94(4), pp.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, pp.24824-24837.



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 ChatGPT-4 and prompt it to issue performance rankings for the ten contest stocks over the following week.

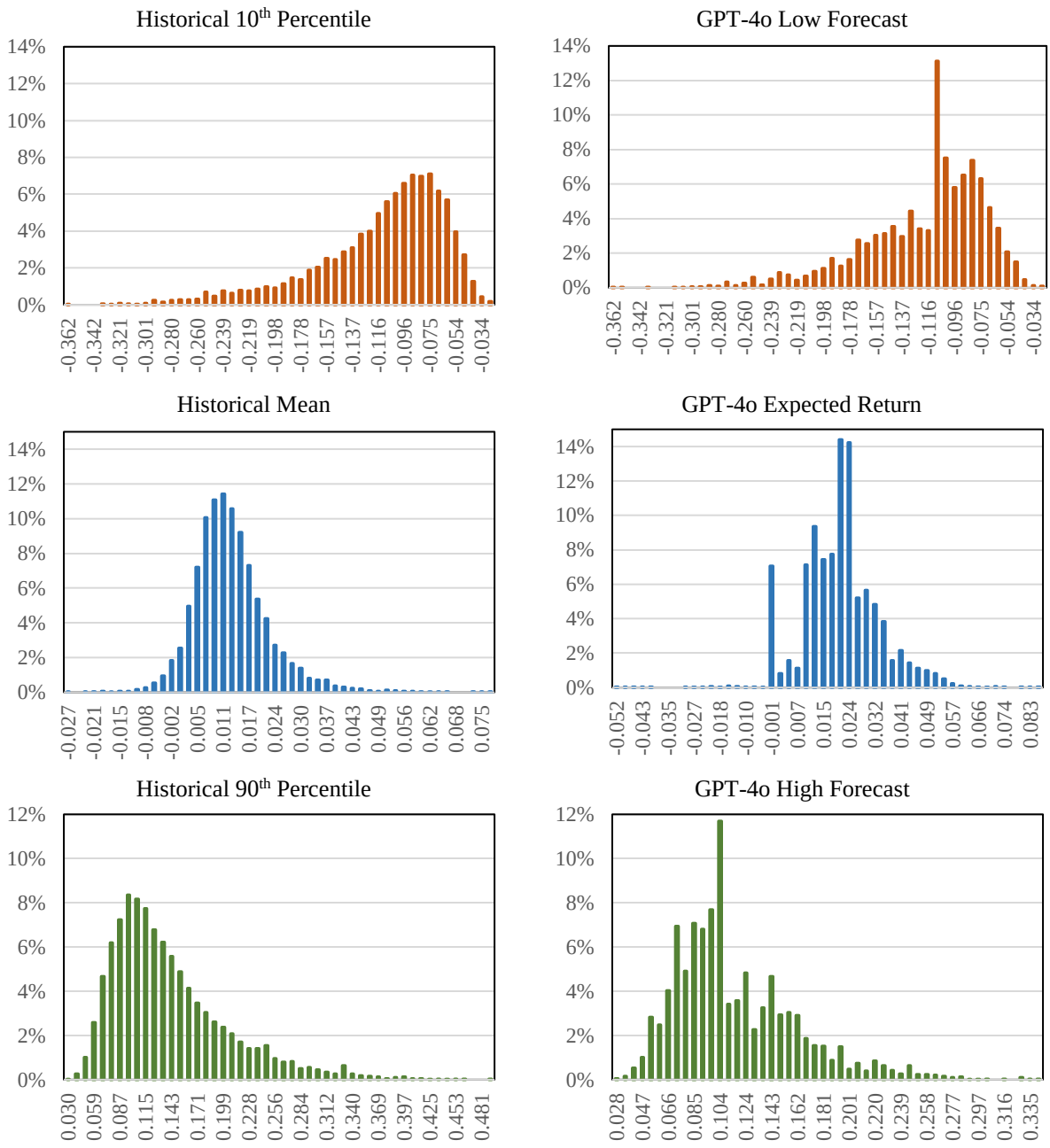


Figure 2. Historical and LLM Forecasts of Low, Expected, and High returns. ChatGPT-4 is provided with up to ten (no fewer than five) years of historical monthly returns for a randomly chosen stock-month, and the process is repeated 10,000 times. The left plots show the distribution of the 10th percentile, mean, and 90th percentile of the historical samples provided to ChatGPT-4. The right plots show the distribution of the resulting next-month 10th percentile, expected, and 90th percentile return forecasts produced by ChatGPT.

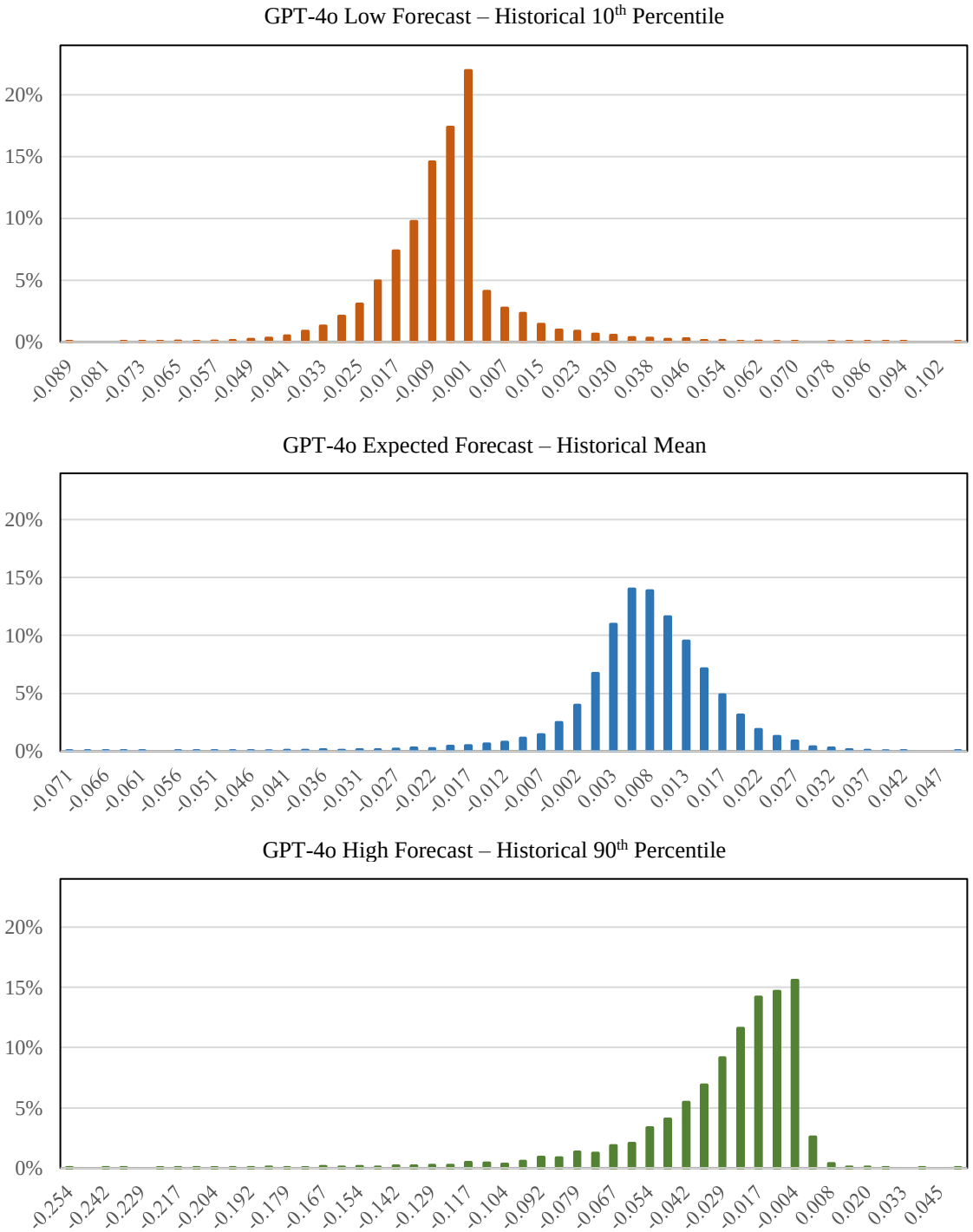


Figure 3. Differences between Historical and ChatGPT Forecasts. ChatGPT-4 is provided with up to ten (no fewer than five) years of historical monthly returns for a randomly chosen stock-month, and the process is repeated 10,000 times. The top (lower) plot shows the distribution of the differences between the 10th (90th) percentile of the historical sample provided to ChatGPT-4 and the resulting next-month 10th (90th) percentile forecast produced by ChatGPT. The middle panel plots the distribution of the differences between the historical mean and ChatGPT's expected return forecast.

Table 1. Sample Statistics

The table presents sample descriptive statistics. Panel A provides contest-stock level statistics for the sample of 1,286 Forcerank contests that occurred between February 2016 and December 2017. Panel B provides statistics for monthly observations from the American Association of Individual Investor survey sample, covering July 1987 through June 2024. In each panel we also include statistics for ChatGPT-4 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
ChatGPT-12week	12,719	5.54	2.85	3.00	6.00	8.00
ChatGPT-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, ChatGPT, 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 ChatGPT-4 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.

	$\text{Forecerank}_{i,t+1}$		$\text{ChatGPT Rank}_{i,t+1}^{12w}$		$\text{Return}_{i,t+1}$
	(1)	(2)	(3)	(4)	(5)
$\text{Forecerank}_{i,t}$			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 ChatGPT-4 based on past 12 weekly returns, and in Specification (3), the stock ranking generated by ChatGPT-4 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.

	$\text{Forecerank}_{i,t+1}$		$\text{ChatGPT Rank}^{12w}_{i,t+1}$		$\text{ChatGPT Rank}^{24w}_{i,t+1}$	
	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 ChatGPT-4 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.

	$\text{Forecerank}_{i,t+1}$		$\text{ChatGPT Rank}_{i,t+1}^{12w}$		$\text{Ret Rank}_{i,t+1}$
	(1)	(2)	(3)	(4)	(5)
$\text{Forecerank}_{i,t}$			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. 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 ChatGPT-4 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.

	ChatGPT $\text{Rank}_{i,t+1}^{12w\ \text{Chart}}$		
	(1)	(2)	(3)
$\text{Forecerank}_{i,t}$		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 6. Extrapolative beliefs: Exponential decay model.

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

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

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 ChatGPT-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 l_1 and a lower l_2 jointly lead to a higher degree of extrapolation and $l_1(1-l_2)$ represents the degree of extrapolation. The sample period is from February 2016 to December 2017.

	$\text{Forecerank}_{i,t}$	$\text{ChatGPT Rank}_{i,t}^{12w}$	$\text{ChatGPT Rank}_{i,t}^{24w}$
	(1)	(2)	(3)
l_1	16.98***	40.72***	45.68***
	(15.53)	(48.89)	(45.30)
l_2	0.28***	0.07***	0.27***
	(6.78)	(5.16)	(18.55)
$l_1(1-l_2)$	12.19	38.03	33.21

Table 7. ChatGPT 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 ChatGPT-4 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 ChatGPT rank. Panel A focuses on the ranking produced by ChatGPT-4o using 12 weekly return lags, and Panel B focuses on ChatGPT-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)
ChatGPT $\text{Rank}_{i,t}^{12w}$	-0.12			-0.41		
	(-0.35)			(-1.15)		
Predicted ChatGPT $\text{Rank}_{i,t}^{12w}$		-0.38			-1.08*	
		(-0.76)			(-1.92)	
Residual ChatGPT $\text{Rank}_{i,t}^{12w}$			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)
ChatGPT $\text{Rank}_{i,t}^{24w}$	-0.50			-0.78**		
	(-1.40)			(-2.05)		
Predicted ChatGPT $\text{Rank}_{i,t}^{24w}$		-0.69			-1.34**	
		(-1.25)			(-2.20)	
Residual ChatGPT $\text{Rank}_{i,t}^{24w}$			-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 8. Market Return Extrapolation: AAII and ChatGPT 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 "bearish" investors minus the percentage of "bullish" investors in the last week of each month. In Specifications (2) to (4), the dependent variable is the ChatGPT-4 sentiment generated based on the 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}	$\text{ChatGPT Sentiment}_{t+1}$			$\text{S\&P Ret}_{t+1,t+6}$
	(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 ChatGPT-4 sentiment measure 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 with "Think step by step when creating your response." In Specification (2), the prompt is followed with "Analyze the data and consider an appropriate model to apply for your response." In Specification (3), the prompt is followed with "Consider human biases that may affect responses to similar questions and avoid these biases when creating your response." In Specification 4, the LLM prompt includes a 500-word summary of Greenwood and Schleifer (2014) obtained from a ChatGPT prompt "Summarize the academic study in 500 words. The text will be used to help guide LLM decision-making." 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.

	ChatGPT Sentiment$_{t+1}$				
	(1)	(2)	(3)	(4)	(5)
S&P Return$_{t}$	9.06***	7.61***	9.65***	5.55***	9.17***
	(16.10)	(11.98)	(17.10)	(7.22)	(16.68)
S&P Return$_{t-1}$	6.41***	6.58***	5.27***	3.40***	5.38***
	(10.65)	(9.81)	(8.78)	(4.52)	(10.20)
S&P Return$_{t-2}$	4.79***	4.42***	3.36***	3.74***	3.99***
	(8.51)	(6.53)	(6.18)	(5.06)	(8.78)
S&P Return$_{t-3}$	3.20***	3.45***	2.67***	2.90***	3.65***
	(5.25)	(5.41)	(5.20)	(3.72)	(7.48)
S&P Return$_{t-4}$	2.33***	2.56***	1.69***	3.27***	2.68***
	(4.06)	(3.98)	(3.13)	(4.05)	(4.89)
S&P Return$_{t-5}$	2.41***	2.59***	1.66***	4.04***	3.39***
	(4.48)	(4.33)	(2.92)	(5.45)	(7.04)
S&P Return$_{t-6}$	3.06***	2.22***	1.91***	1.83**	2.48***
	(6.75)	(3.47)	(3.60)	(2.51)	(6.03)
S&P Return$_{t-7}$	1.71***	1.88***	0.97**	1.92***	2.00***
	(3.59)	(2.85)	(2.15)	(2.72)	(4.43)
S&P Return$_{t-8}$	1.59***	1.44**	2.25***	0.52	1.78***
	(3.21)	(2.42)	(4.67)	(0.70)	(3.69)
S&P Return$_{t-9}$	1.44***	2.53***	1.01**	1.65**	2.13***
	(2.95)	(3.84)	(2.08)	(2.18)	(4.83)
S&P Return$_{t-10}$	0.34	1.71***	1.31***	0.59	1.47***
	(0.73)	(2.93)	(2.89)	(0.78)	(3.26)
S&P Return$_{t-11}$	0.85*	0.64	0.97*	2.06***	2.12***
	(1.79)	(1.25)	(1.95)	(2.95)	(5.11)
Observations	438	438	438	438	438
R-squared	0.630	0.521	0.625	0.295	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 ChatGPT-4. Panel A presents descriptive statistics for historical returns, GPT-4o forecasts, and realized returns. Panel B examines bias tests, comparing the expected forecasts against historical means and realized outcomes. Panel C provides calibration evidence, evaluating the accuracy of the Low and High forecasts relative to their historical percentiles. All statistics are based on 10,000 stock-month observations randomly selected from the 1926 to 2023 period, with up to ten years of historical monthly returns provided to ChatGPT-4.

Panel A: Descriptive statistics

Variable	Obs.	Mean	Std.	5%	25%	50%	75%	95%
Expected forecast	10,000	2.23	1.23	0.00	1.45	2.27	2.84	4.37
Historical mean	10,000	1.40	0.94	0.14	0.81	1.29	1.85	3.07
Realized returns	9954	1.15	14.61	-15.69	-5.42	0.00	5.71	20.34
Low forecast	10,000	-11.50	4.66	-20.78	-13.97	-10.53	-7.97	-5.64
Historical 10%	10,000	-11.02	5.21	-22.02	-13.50	-9.76	-7.26	-4.83
High forecast	10,000	11.89	4.55	6.20	8.68	10.75	14.54	20.83
Historical 90%	10,000	14.48	6.36	7.02	9.98	12.95	17.46	27.27
Confidence interval %	10,000	23.39	8.87	12.23	17.04	21.29	28.46	41.20
Historical 90% - 10%	10,000	25.50	11.25	12.11	17.43	22.67	30.88	48.87

Panel B: Forecast bias

Difference	Mean Difference	p-Value
Expected forecast = Historical mean	0.83	0.000
Expected forecast = Realized return	1.08	0.000
Low forecast = Historical 10%	-0.49	0.000
High forecast = Historical 90%	-2.60	0.000

Panel C: Realized returns relative to historical and ChatGPT forecasts

% of realized returns below low forecast	10.33
% of realized returns in confidence interval	76.53
% of realized returns above high forecast	12.64
% of realized returns below historical 10%	11.61
% of realized returns in historical interval	79.02
% of realized returns above historical 90%	9.30

Table 11. ChatGPT Return Forecasts: The Role of Historical Return Magnitude

This table presents the results from linear regressions conducted at the stock-month level, as specified in Equation (5) of the main text. The stock-month sample is constructed by randomly selecting 100 months between 1926 and 2023 and then selecting 10 stocks from each size decile, based on Fama-French size breakpoints. In Column (1), the dependent variable is the 10th percentile return forecast generated by ChatGPT-4, based on individual stock returns over the past 120 months. Columns (2) and (3) use the mean return forecast and the 90th percentile return forecast generated by ChatGPT-4 as the dependent variables, respectively. The explanatory variables are the minimum, 10^{th} to 90^{th} percentiles, and maximum of the 120 realized monthly returns for each stock. All specifications include year-month fixed effects. Standard errors are clustered by year-month. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Historical Returns	ChatGPT $\text{Low}_{i,t+1}$ (1)	ChatGPT $\text{Exp}_{i,t+1}$ (2)	ChatGPT $\text{High}_{i,t+1}$ (3)
Minimum Return	0.01***	0.01***	0.00
	(6.30)	(26.23)	(1.66)
10^{th} Percentile	0.81***	0.10***	-0.14***
	(52.88)	(27.36)	(-8.83)
20^{th} Percentile	-0.05**	0.09***	0.01
	(-1.99)	(17.71)	(0.28)
30^{th} Percentile	-0.01	0.08***	-0.04
	(-0.46)	(15.01)	(-1.08)
40^{th} Percentile	0.00	0.09***	-0.02
	(0.08)	(11.44)	(-0.48)
50^{th} Percentile	0.03	0.10***	-0.03
	(0.98)	(12.24)	(-0.76)
60^{th} Percentile	-0.02	0.08***	0.04
	(-0.92)	(9.12)	(0.99)
70^{th} Percentile	-0.02	0.08***	-0.05
	(-0.80)	(13.26)	(-1.25)
80^{th} Percentile	-0.05***	0.10***	0.49***
	(-3.44)	(21.19)	(20.54)
90^{th} Percentile	-0.04***	0.13***	0.32***
	(-4.66)	(51.27)	(18.49)
Maximum Return	0.00***	0.01***	-0.00
	(4.00)	(37.05)	(-0.13)
Year-Month Fixed Effects	Yes	Yes	Yes
Observations	10,000	10,000	10,000
R-squared	0.937	0.907	0.864

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 ChatGPT-4 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.

	$\text{Forecerank}_{i,t+1}$		$\text{ChatGPT Rank}_{i,t+1}^{12w}$		$\text{Adj Return}_{i,t+1}$
	(1)	(2)	(3)	(4)	(5)
$\text{Forecerank}_{i,t}$			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, ChatGPT, 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 ChatGPT-4 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.

	$\text{Forecerank}_{i,t+1}$		$\text{ChatGPT}^{24w}_{i,t+1}$		$\text{Return}_{i,t+1}$
	(1)	(2)	(3)	(4)	(5)
$\text{Forecerank}_{i,t}$			0.23***	0.13***	
			(23.63)	(16.19)	
$\text{Return}_{i,t}$	12.73***	33.69***		32.04***	1.05
	(20.26)	(37.70)		(36.76)	(0.73)
$\text{Return}_{i,t-1}$	2.39***	9.66***		9.35***	-2.67**
	(4.47)	(14.75)		(14.49)	(-2.03)
$\text{Return}_{i,t-2}$	2.22***	2.53***		2.25***	-3.79***
	(4.25)	(4.58)		(4.20)	(-2.69)
$\text{Return}_{i,t-3}$	2.52***	2.07***		1.74***	-2.64*
	(4.74)	(3.70)		(3.22)	(-1.87)
$\text{Return}_{i,t-4}$	2.31***	4.34***		4.04***	0.64
	(4.32)	(6.99)		(6.67)	(0.46)
$\text{Return}_{i,t-5}$	2.35***	4.41***		4.11***	2.69**
	(4.44)	(7.65)		(7.32)	(1.99)
$\text{Return}_{i,t-6}$	1.59***	0.59		0.39	5.45***
	(3.09)	(1.03)		(0.70)	(3.63)
$\text{Return}_{i,t-7}$	1.41***	1.06*		0.88	-1.21
	(2.69)	(1.78)		(1.52)	(-0.88)
$\text{Return}_{i,t-8}$	1.32***	1.10**		0.93*	-2.11
	(2.58)	(1.97)		(1.73)	(-1.50)
$\text{Return}_{i,t-9}$	0.99**	0.60		0.47	-3.02**
	(1.98)	(1.08)		(0.87)	(-2.17)
$\text{Return}_{i,t-10}$	1.83***	1.74***		1.50***	-0.61
	(3.71)	(3.07)		(2.72)	(-0.43)
$\text{Return}_{i,t-11}$	0.20	0.89*		0.86*	-2.00
	(0.43)	(1.69)		(1.68)	(-1.51)
$\text{Return}_{i,t-12}$	0.80*	0.61		0.51	-1.08
	(1.67)	(1.17)		(1.00)	(-0.74)
$\text{Return}_{i,t-13}$	0.99**	1.05**		0.93*	-0.13
	(2.20)	(2.00)		(1.80)	(-0.09)
$\text{Return}_{i,t-14}$	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 fore 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 of 2024, and the adjacent in-training-sample period is March to December of 2023. As in Table 2, ChatGPT-4 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.

	ChatGPT $\text{Rank}_{i,t+1}^{12w}$		
	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 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 ChatGPT-4 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.

	ChatGPT $\text{Rank}_{i,t+1}^{24w\ \text{Chart}}$		
	(1)	(2)	(3)
$\text{Forecerank}_{i,t}$		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 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 included 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.

	ChatGPT $\text{Rank}_{i,t+1}^{12w}$		
	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 Summary of Greenwood and Shleifer (2014)

Prompt:

Summarize the [included] academic study in 500 words. The text will be used to help guide LLM decision-making.

Output:

The study "Expectations of Returns and Expected Returns" by Robin Greenwood and Andrei Shleifer explores investor expectations of stock market returns and contrasts these with model-based expected returns (ERs) derived from economic data. Using data spanning 1963–2011, sourced from surveys such as Gallup, the Graham-Harvey CFO Survey, and others, the paper uncovers significant insights into the nature of investor expectations.

Key Findings:

1. **Positive Correlation Across Surveys**: Expectations of returns are consistent across the six data sources. These expectations exhibit a strong relationship with past stock returns and the level of the stock market, aligning with a pattern of extrapolative behavior by investors.
2. **Negative Correlation with Model-Based ERs**: Contrary to rational expectations theories, when investors anticipate high returns, model-based ERs tend to be low. This finding challenges the idea that investor beliefs align with the theoretical rational expectations of a representative investor.
3. **Predictive Power and Opposite Signs**: Both investor expectations and ERs predict future market returns but in opposite directions. High investor expectations are associated with low future returns, while high ERs predict high future returns.

Methodology:

The researchers use survey data as proxies for investor expectations, comparing them to ERs derived from aggregate variables like dividend-price ratios and consumption-based measures. Statistical tests reveal a robust link between expectations and past returns, with investors strongly influenced by recent performance and current market levels.

Implications:

- The results suggest that survey-based expectations reflect actual investor beliefs, which are shaped by extrapolation from recent trends rather than rational assessments of fundamental values.
- These findings question the validity of standard financial models that rely on rational expectations and indicate the importance of behavioral factors in understanding market dynamics.

Relevance for Decision-Making:

This study highlights how investor expectations, driven by extrapolation, can diverge from fundamental-based predictions. Large language models (LLMs) can incorporate these insights to better interpret and predict market sentiment by considering behavioral biases and historical data relationships. This approach aligns with improving LLM-driven financial decision-making, as it underscores the predictive and psychological factors influencing markets.

Table IA7. Market Return Extrapolation: Claude Sentiment

This table presents the results from linear regressions at the month level. In Specification (1), the dependent variable is Claude sentiment generated based on the US stock market (S&P 500) returns in the past 12 months (Specifications 1-3) or 36 months (Specification 4). The independent variables include AAII sentiment, measured as the percentage of "bearish" investors minus the percentage of "bullish" investors in the last week of each month, and ChatGPT sentiment. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample period is from July 1987 to June 2024.

		Claude Sentiment$_{t+1}$ 12 months		Claude Sent$_{t+1}$ 36 months
	(1)	(2)	(3)	(4)
AAII Sent$_t$		1.51***		
		(7.53)		
ChatGPT-4o Sent$_t$			0.81***	
			(23.90)	
S&P Return$_t$	10.64***			4.67***
	(17.74)			(7.03)
S&P Return$_{t-1}$	5.74***			4.71***
	(9.83)			(8.03)
S&P Return$_{t-2}$	3.60***			2.15***
	(6.52)			(4.00)
S&P Return$_{t-3}$	2.61***			2.56***
	(4.85)			(4.57)
S&P Return$_{t-4}$	1.92***			1.83***
	(3.34)			(3.49)
S&P Return$_{t-5}$	2.31***			1.85***
	(3.95)			(3.50)
S&P Return$_{t-6}$	1.29**			1.61***
	(2.30)			(3.06)
S&P Return$_{t-7}$	1.43***			1.62***
	(2.87)			(3.27)
S&P Return$_{t-8}$	1.43***			1.58***
	(2.63)			(3.07)
S&P Return$_{t-9}$	1.05**			1.09**
	(2.00)			(2.14)
S&P Return$_{t-10}$	1.17**			1.05**
	(2.49)			(2.17)
S&P Return$_{t-11}$	0.66			0.58
	(1.37)			(1.30)
Observations	438	438	438	435
R-squared	0.630	0.133	0.631	0.453

Table IA8. Market Return Extrapolation: ChatGPT-o1 Sentiment

This table presents the results from linear regressions at the month level. In Specification (1), the dependent variable is ChatGPT-o1 sentiment generated based on the US stock market (S&P 500) returns in the past 12 months (Specifications 1-3) or 36 months (Specification 4). The independent variables include AAII sentiment, measured as the percentage of "bearish" investors minus the percentage of "bullish" investors in the last week of each month, and ChatGPT-4 sentiment. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample period is from July 1987 to June 2024.

		ChatGPT-o1 Sent_{t+1} 12 months		ChatGPT-o1 Sent_{t+1} 36 months
	(1)	(2)	(3)	(4)
AAII Sent_t		1.38***		
		(6.30)		
ChatGPT-4o Sent_t			0.80***	
			(24.00)	
S&P Return_t	8.94***			8.41***
	(13.30)			(12.41)
S&P Return_{t-1}	7.65***			6.21***
	(10.78)			(8.35)
S&P Return_{t-2}	4.64***			4.83***
	(7.46)			(6.41)
S&P Return_{t-3}	2.27***			2.76***
	(3.59)			(3.99)
S&P Return_{t-4}	2.49***			2.74***
	(3.99)			(4.20)
S&P Return_{t-5}	2.32***			4.13***
	(3.92)			(6.28)
S&P Return_{t-6}	1.65***			2.05***
	(2.61)			(2.99)
S&P Return_{t-7}	-0.44			0.82
	(-0.76)			(1.43)
S&P Return_{t-8}	1.42**			1.21**
	(2.27)			(2.06)
S&P Return_{t-9}	0.78			0.22
	(1.22)			(0.38)
S&P Return_{t-10}	0.31			0.90
	(0.46)			(1.48)
S&P Return_{t-11}	0.41			0.25
	(0.73)			(0.45)
Observations	438	438	438	438
R-squared	0.534	0.095	0.503	0.516

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 ChatGPT-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.

	ChatGPT 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
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)
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)
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)
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)
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)
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)
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)
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)
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)
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)
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)
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