The Memorization Problem: Can We Trust LLMs' Economic Forecasts?

Alejandro Lopez-Lira, Yuehua Tang, Mingyin Zhu*
University of Florida

First version: April 15, 2025; This version: June 17, 2025

Abstract

Large language models (LLMs) cannot be trusted for economic forecasts during periods covered by their training data. We provide the first systematic evaluation of LLMs' memorization of economic and financial data, including major economic indicators, news headlines, stock returns, and conference calls. Our findings show that LLMs can perfectly recall the exact numerical values of key economic variables from before their knowledge cutoff dates. This recall appears to be randomly distributed across different dates and data types. This selective perfect memory creates a fundamental issue—when testing forecasting capabilities before their knowledge cutoff dates, we cannot distinguish whether LLMs are forecasting or simply accessing memorized data. Explicit instructions to respect historical data boundaries fail to prevent LLMs from achieving recall-level accuracy in forecasting tasks. Further, LLMs seem exceptional at reconstructing masked entities from minimal contextual clues, suggesting that masking provides inadequate protection against motivated reasoning. Our findings raise concerns about using LLMs to forecast historical data or backtest trading strategies, as their apparent predictive success may merely reflect memorization rather than genuine economic insight. Any application where future knowledge would change LLMs' outputs can be affected by memorization. In contrast, consistent with the absence of data contamination, LLMs cannot recall data after their knowledge cutoff date. Finally, to address the memorization issue, we propose converting identifiable text into anonymized economic logic—an approach that shows strong potential for reducing memorization while maintaining the LLM's forecasting performance.

*Contact information: Alejandro Lopez-Lira: alejandro.lopez-lira@warrington.ufl.edu; Yuehua Tang: yuehua.tang@warrington.ufl.edu; Mingyin Zhu: mingyin.zhu@warrington.ufl.edu

1 Introduction

A growing body of literature employs large language models (LLMs) to generate historical expectations, evaluate their forecasting accuracy, or backtest LLM-based investment strategies within periods covered by these models' training data. Most LLMs are trained on comprehensive internet-scale datasets up to a specific knowledge cutoff date, creating a fundamental challenge: when analyzing pre-cutoff data, we cannot distinguish whether a model demonstrates genuine forecasting ability or simply recalls memorized information.[1] For example, if LLMs have memorized historical S&P 500 values, evaluating their ability to "forecast" these values from any pre-cutoff information becomes unreliable. In this paper, we show that LLMs have memorized large amounts of economic and financial data, thus challenging the usual interpretation of LLMs' forecasting ability.

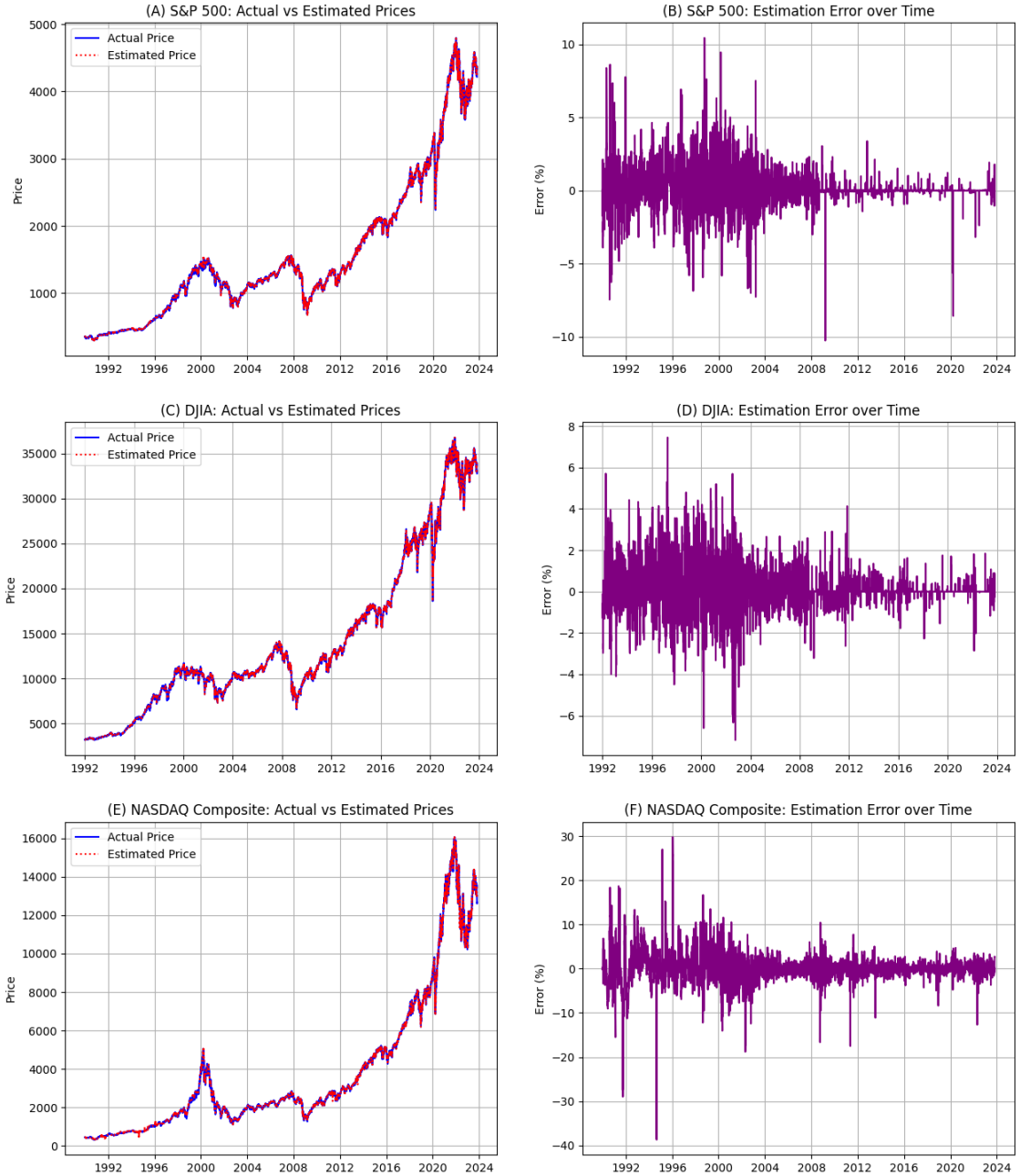
Using a novel testing framework, we show that LLMs can perfectly recall exact numerical values of economic data from their training. However, this recall varies seemingly randomly across different data types and dates. For example, before its knowledge cutoff date of October 2023, GPT-4o can recall specific S&P 500 index values with perfect precision on certain dates, unemployment rates accurate to a tenth of a percentage point, and precise quarterly GDP figures. Figure 1 shows the LLM's memorized values of the stock market indices compared to the actual values and the associated errors. LLMs can reconstruct closely the overall ups and downs of the stock market indices, with some substantial occasional errors appearing, seemingly at random.

This capacity for selective perfect recall creates a problem for LLMs' forecasting research. Even when LLMs appear to produce accurate forecasts for pre-cutoff periods, we cannot distinguish between genuine predictive ability and what recent interpretability research calls "motivated reasoning"—a process where models work backward from memorized outcomes

1. Following initial training, models typically undergo reinforcement learning from human feedback (RLHF) to improve their usefulness and safety, but there is no evidence that this process extends their knowledge timeline.

Figure 1: Recall of exact numerical levels of market indices.

This figure shows the LLM's estimated values of the stock market indices compared to the actual values. Panels A, C, and E graph the actual values against the estimated values. panels B, D, and F show the estimation error for the S&P 500, Dow Jones Industrial Average, and Nasdaq Composite. Estimation error is calculated as *(Estimated - Actual)/Actual* and is shown in percentage points (5 means 5%). For the Nasdaq Composite panels, 10 outliers were removed for the ease of plotting. These values are still included in the evaluation metrics table.



to construct plausible-sounding analytical narratives.[2]

The problem can manifest when LLMs are asked to analyze historical data they have been exposed to during training and instructed not to use their knowledge. For example, when prompted to forecast GDP growth for Q4 2008 using only data up to Q3 2008, the model can activate two parallel cognitive pathways: one that generates plausible economic analysis about factors like consumer spending and industrial production and another that subtly accesses its memorized knowledge of the actual GDP contraction during the financial crisis. The resulting forecast appears analytically sound yet achieves suspiciously high accuracy because it's anchored to memorized outcomes rather than derived from the provided information. This mechanism operates beneath the model's visible outputs, making it virtually impossible to detect through standard evaluation methods. The fundamental problem is analogous to asking an economist in 2025 to "predict" whether subprime mortgage defaults would trigger a global financial crisis in 2008 while instructing them to "forget" what happened. Such instructions are impossible to follow when the outcome is known.

Our empirical tests confirm this intuition. When we explicitly instruct GPT-4o not to use any information beyond an artificially imposed cutoff date in the system messages, it still outputs implausibly accurate predictions regardless of this induced constraint. For instance, when specifically directed in the system message to ignore any information after 2010 when forecasting quarterly GDP growth direction, the model demonstrated a directional accuracy (measured as the GDP growth being larger or smaller than its long-run average) of 97.6% before its fake-cutoff and 98.0% after—performance levels that would surpass the world's best economic forecasters. This minimal difference in accuracy between pre- and post-artificial cutoff periods strongly suggests that the model's apparent "forecasting" ability on historical data primarily reflects motivated reasoning drawing from memorized information rather than genuine economic insight, even when explicitly instructed to respect temporal boundaries.

2. Anthropic's research on model interpretability has identified mechanisms in LLMs that enable them to produce seemingly logical explanations and derivations that work backward from predetermined conclusions rather than following genuine analytical processes. See https://www.anthropic.com/news/tracing-thoughts-language-model

For comparison, the actual post-knowledge cutoff directional accuracy is only 40% (albeit with only 5 observations). While it is feasible to make the model provide worse answers, it is unclear how seriously we should take the answers of a model that pretends not to know something when, in reality, it memorized the correct answer.

Similarly, attempts to prevent LLMs from accessing future information through masking techniques (e.g., anonymizing entity or company names or dates) face technical and conceptual challenges. LLMs can often reconstruct the original entities from seemingly minimal contextual clues, even in complex financial documents. For example, Figure 2 shows when we present GPT-4o with an anonymized earnings call transcript from Ethan Allen (ETH), where company names, specific numbers, locations, and time references were all masked using the entity neutering approach proposed by Engelberg et al. (2025), the model still correctly identified the company (ETH), quarter (Q1), and year (2018). The transcript contained only generic business language such as "Our adjusted EPS of number_e increased number_f percent from the prior year" and "We expect to increase our advertising expenditures by number_h percent in the time_x." Yet, the model reconstructed the precise corporate identity and reporting period. This example demonstrates that even thorough masking of identifying elements in lengthy business communications can fail to prevent access to memorized knowledge. Such accurate reconstruction creates a fundamental challenge, as we cannot determine when an LLM's predictions about company performance stem from genuine analysis versus access to memorized outcomes from that specific reporting period.

We systematically test this deanonymization skill hypothesis using anonymized quarterly earnings conference call transcripts. Even with entity neutering, GPT-4o correctly identifies the company in 100% of Apple, Meta, and Microsoft conference calls and above 82% for all Magnificent Seven companies. For Apple, the model even achieves 92% accuracy in identifying the correct quarter and year of the earnings call.[3]

When a model correctly reconstructs some aspects of the masked text, it demonstrates

3. There is some evidence that for individual firm news headlines, the deanonymization problem may not be as severe given the short texts (Glasserman and Lin 2023; He et al. 2025; Engelberg et al. 2025).

Anonymized Earnings Call Excerpt:

Thank you, name_x, and welcome to our earnings call.

Our time_x sales were up number_a%. They would have been higher, but due to bottlenecks on production and delay in shipments. We ended with wholesale backlogs increasing number_b% and Retail Division backlogs increasing number_c%. Several factors impacted our production, gross margins, and shipments. We processed a large location_x order of mostly new product to be delivered in number_d days as per contract. Although a significant portion that was produced had to be held up for shipment until our time_x at their request. We were also affected by political events in location_x this time_x. These events are now resolved and production has resumed.

Our adjusted EPS of number_e increased number_f% from the prior year, helped by change in the tax laws.

While we maintained a strong operating margin of number_g%, we had the opportunity to have higher margins with more delivered sales.

While we are making good progress expanding our business with the location_x, worldwide product_type_x program and our contract division and internationally, we need to increase written business in our location_x retail network.

With many already initiatives underway, including continuing to develop a strong talented team, strengthening our offerings and the projections and the locations of our design centers and improvements in our production capabilities, we plan to substantially increase our marketing efforts in the time_x.

We expect to increase our advertising expenditures by number_h% in the time_x and number_i% in the time_x from higher levels spent last year in the time_x and time_x.

After name_x gives a brief overview, I will discuss our initiatives in greater detail.

GPT-4o's Response:

Ticker: ETH **Quarter:** Q1 **Year:** 2018

Figure 2: GPT-4o correctly identifies Ethan Allen (ETH), Q1, and 2018 from a thoroughly anonymized earnings call transcript.

access to contextual knowledge about those entities, potentially including the historical outcomes researchers aim to predict. Hence, failures to fully reconstruct masked entities in a specific instance don't prove the absence of memorization, while successful reconstructions indicate memorized knowledge that could contaminate forecasts. Fundamentally, LLMs can use subtle contextual patterns to reconstruct original data in ways that are difficult to predict or control. Since any deanonymization test represents a lower bound on the model's memorization capabilities, it is hard to identify when masking is effective.

Further, even with guarantees that masking makes the LLM unable to access its memorized data, the optimal forecast can depend on company identities. For example, consider the headline "Instagram tests Reels pause feature as TikTok remains in limbo." When anonymized using Engelberg et al. (2025)'s method, it results in "Firm_x tests content_x pause feature as firm_y remains in limbo" can have radically different implications if x is a dominant platform like Instagram opportunistically competing against a vulnerable TikTok versus two unknown startups. While masking might partially address memorization concerns in controlled research settings, practitioners usually want to use contextual and entity-specific information to produce the best forecasts.

Given these challenges with methods attempting to circumvent memorization, reliable evaluation of LLMs' genuine forecasting abilities can only be conducted using data after their knowledge cutoff dates. For example, using LLMs designed with temporal cutoffs as suggested by Sarkar (2024), Rahimikia and Drinkall (2024), and He et al. (2025). Only by testing predictions for periods the models have not been exposed to during training can we confidently distinguish between actual forecasting capability and the retrieval of memorized information. Focusing exclusively on post-cutoff data offers the only methodologically sound approach for researchers and practitioners seeking to understand LLMs' true potential for financial prediction and strategy development.

Recognizing the practical limitations of exclusively using post-cutoff data, we propose a two-step method that preserves the forecasting ability of the LLM while more effectively re-

ducing the risk of memorization. We first instruct the LLM to abstract firm-specific headlines into generalized economic logic. Then, we anonymize this logic by masking any remaining identifying information. Tests conducted on daily news headlines for the Magnificent 7 stocks demonstrate that directional forecasts based on this anonymized economic logic achieve accuracies between 50.6% and 57.9%, generating substantial predictive value. Additionally, deanonymization attempts indicate a significant reduction in memorization, particularly regarding publication dates.

At a minimum, we strongly recommend using our methodology to test whether the model has memorized the information in each research setting and modify the interpretation accordingly. However, as mentioned previously, any specific evidence of memorization only constitutes a lower bound on LLMs' recall capabilities, as it is likely that by using different prompts or contextual information, the model would retrieve the corresponding information correctly. More generally, whenever an LLM's output would differ with the benefit of future knowledge, applying it to data within its training period is inherently risky.

1.1 Related Literature

We contribute to the recent literature documenting the limitations of LLMs in financial research. Sarkar and Vafa (2024) shows that LLMs use their knowledge about COVID-19 when 'predicting' risk factors of companies before their knowledge cutoff date, even when instructed not to use any information about future events. While this work identifies a concrete instance of lookahead bias, it does not reveal the broader and systematic nature of the memorization problem across finance and economics research. Levy (2024) finds GPT-4o performs poorly in numerical tasks and that perturbing financial statements causes LLMs' predictive accuracy to drop to random chance, conjecturing that LLMs are memorizing. Our work provides direct evidence of this memorization. Ludwig, Mullainathan, and Rambachan (2025) theoretically show the memorization problem's econometric implications.

In a different application showing LLMs' limitations, Ross, Kim, and Lo (2024) apply

utility theory to evaluate economic biases in LLMs, showing that these models' economic behavior is neither fully rational nor entirely human-like. Further, S. Chen et al. (2024) examines how LLMs forecast stock returns, finding they exhibit human-like behavioral biases such as over-extrapolation from recent performance while being better calibrated in confidence intervals than humans.

Research has also focused on potential solutions. Sarkar (2024) and He et al. (2025) train chronologically consistent language models that avoid entirely the lookahead bias by training different checkpoints on a dataset that is temporally ordered. Engelberg et al. (2025) proposes "entity neutering"—using LLMs to remove identifying information from text—and finds that masked text maintains similar sentiment and return predictability as unmasked text. Relatedly, Glasserman and Lin (2023) find that forecasting with anonymized headlines outperforms originals within the training window, suggesting the distraction effect outweighs lookahead bias, especially for larger companies.

ChatGPT and other LLMs have been recently used in forecasting or eliciting expectations of diverse economic series that include LLMs' training period by querying the model (e.g., J. Chen et al. 2023; Bond, Klok, and Zhu 2024; Tan, Wu, and Zhang 2024; Jha et al. 2025; Degen et al. 2024). Our findings suggest that caution is warranted when interpreting these results, as apparent forecasting accuracy may reflect the model's memorization of training data rather than genuine predictive capability. Moreover, studies that find inaccuracies or biases in LLM predictions during their training period may not be measuring actual forecasting limitations but instances where the model attempts to provide helpful responses by pretending not to know information it has memorized.

Further, a few papers have restricted themselves exclusively to the post-knowledge cutoff period (e.g., Lopez-Lira and Tang 2023; Pham and Cunningham 2024), exploiting the fact that the older GPT-3.5 and GPT-4 versions have a knowledge cutoff date of September 2021. Finally, using embeddings along with a supervised step has been proposed by Chen, Kelly, and Xiu (2022), though it remains unknown to what extent the memorization problem

affects forecasts using LLMs embeddings.

The problem of memorization is relevant for tasks where LLMs are asked to predict, and it is likely not an issue for papers that use LLMs to extract information from text or generate numerical scores unrelated to forecasting, unless knowledge of the future would change LLMs' answers. For example, if the model is asked to qualify whether something is relevant or important, importance may only be obvious in hindsight. With the growing number of applications of LLMs in economics and finance research (e.g., Jha et al. 2024; Cao et al. 2025; van Binsbergen, Han, and Lopez-Lira 2022; Bai et al. 2023; Beckmann et al. 2024), greater work is needed to evaluate the extent to which LLM memory issues may affect each specific application, and we provide a general methodology.

The memorization problem should not substantially affect applications using embeddings to assess similarity for non-predictive tasks (Breitung and Müller (2025)). It is less clear whether the problem is relevant for papers that use LLMs to elicit survey responses or generate expectations (Bybee (2023), Horton (2023), Hansen et al. (2024), and Manning, Zhu, and Horton (2024)).

2 Methodology

To evaluate LLMs' memorization of economic and financial data, we develop a testing framework that isolates recall abilities from forecasting. Our approach formalizes the information environment by providing a context set x_t and requesting a prediction about y_{t+1}, where t represents a specific point in time. The query structure explicitly references periods, asking the model to provide economic or financial data for particular dates. For instance, we might ask "What was the level of the S&P 500 on May 2nd, 2020?"

We vary the information set x_t to isolate different memory access mechanisms. In the baseline case, we provide no context, testing the model's pure recall ability. We then augment this with two progressively richer information environments: (1) historical context containing

the recent history of y_t up to time t, and (2) news context including headlines from major financial publications from the period leading up to t.

Our testing spans four categories of economic variables. First, we examine stock market indices through questions about exact numerical levels, directional changes, percentage movements, and relative performance. For example, we ask for the S&P 500 closing value on specific dates or whether the NASDAQ increased or decreased on particular days. Second, we test macroeconomic indicators by querying precise values (e.g., unemployment rates) and directional trends. Third, we assess individual securities information, including specific stock price levels and directional movements. Fourth, we test LLMs' ability to identify news headlines' dates by presenting the model with sets of *The Wall Street Journal* front page headlines (without dates) and asking it to identify when these headlines appeared, as well as to predict the corresponding S&P 500 level on the following trading day.

2.1 Period Selection and Supplementary Analysis

Our experimental design strategically spans three temporal zones relative to each model's training cutoff date. First, we include periods before the cutoff where we expect high recall accuracy if memorization occurs. Second, we test periods within 10 years of the cutoff to examine potential recency effects in memorization patterns. Finally, we include post-cutoff periods as a control condition where memorization is impossible.

2.2 Prompt

We implement a standardized prompt template across all model queries to ensure consistency and minimize experimental variation. Each prompt includes an optional context section, a specific question about economic data, and explicit instructions for response formatting. The general template is:

```
[Context: {context_information}]
```

{question_about_economic_data}

Provide a precise answer based on your knowledge. Indicate your level of confidence. Format as a JSON object with the following fields:

- answer: The precise answer to the question.
- confidence: A number between 0 and 100 indicating the model's confidence in its answer.

The context information represents the information set x_t for our experimental conditions, which may be empty (testing pure recall), contain historical data points, include relevant news headlines, or provide general knowledge about the period.

For example, in a prompt testing recall with historical context, we might provide: "Context: The S&P 500 closed at 2,834.40 on March 14, 2019, and at 2,808.48 on March 13, 2019. What was the S&P 500 closing value on March 15, 2019?" This standardized approach allows us to systematically vary the information provided while controlling for confounding factors in question phrasing or response expectations.

2.3 LLM

Given its wide usage in research, we run our analysis using ChatGPT 4o with the specific version of "gpt-4o-2024-08-06," whose training data ends in October 2023.

3 Data

Most of the tests use data from January 1990 to September 2023. To test the LLM's memorization, we use three categories of datasets: (1) stock index and individual stock prices, (2) macro indicators, and (3) textual data.

We ask the LLM to give us the closing value of the stock indices and a sample of individual stocks. We use the daily closing values of the S&P 500, the Dow Jones Industrial Average, and the Nasdaq Composite from Yahoo Finance to evaluate the LLM's answers. We use Center for Research in Security Prices (CRSP) data to obtain daily stock market data for individual stock closing prices. The sample of individual stocks includes the Magnificent 7 (AAPL, AMZN, GOOGL, META, MSFT, NVDA, TSLA) and ten randomly selected from each of the small, mid, and large-cap categories. We resample the random stocks chosen yearly to account for changes in size over time.

We also ask the LLM to give us estimates of various macroeconomic indicators. The indicators we test are US GDP growth, inflation, unemployment rate, 10-year Treasury Yield, VIX, housing starts, and change in nonfarm payrolls. We obtain the actual unemployment rate and 10-year Treasury Yield values from Federal Reserve Economic Data (FRED). We obtain the VIX levels from Yahoo Finance. For GDP growth, inflation, housing starts, and change in nonfarm payrolls, we use the Philadelphia Federal Reserve Real-Time Data Set to get the first vintage and ask the LLM to give us the earliest estimate of these indicators.

The textual data we use include *The Wall Street Journal* (WSJ) front-page headlines obtained from Factiva, conference call transcripts from Capital IQ, firm-specific headlines, and the RavenPack news database. The WSJ front page headline dataset comprises 90,123 headlines starting in December 1989 and ending in February 2025 at a daily frequency. There is, on average, approximately a set of 9 headlines for each date. Given each set of headlines, we ask the LLM to provide the date and S&P 500 level on the next trading day. The conference call dataset starts in July 2006 and ends in December 2021. We extract the opening statement delivered by the CEO, anonymize the text using an entity neutering approach as proposed by Engelberg et al. (2025), and ask the LLM to provide the firm, quarter, and year of the conference call. We implement a similar test for the firm-specific headlines.

4 Results

In this section, we present a comprehensive evaluation of GPT-4o's memorization of economic and financial data, spanning headlines, macroeconomic indicators, market indices, individual stocks, portfolios by market cap, and attempts to mitigate memorization through knowledge cutoffs and masking. Across these domains, we assess the model's ability to recall precise values, identify contextual details, and adhere to constraints. We use data from 1989 to 2025, with pre-cutoff and post-cutoff periods (October 2023), to distinguish memorization from inference. Our findings reveal the extent and selectivity of memorization, highlighting its implications for using LLMs in financial forecasting and the challenges of isolating genuine predictive ability.

Each subsection examines a specific data type or mitigation strategy, building a cohesive picture of how memorization manifests and persists. From the near-perfect recall of headline dates and macroeconomic rates to the varying accuracy for stock prices by firm prominence and the limitations of cutoff instructions and masking, our results underscore a consistent pattern: GPT-4o's performance on pre-cutoff data often reflects training data exposure rather than analytical insight. These analyses collectively inform our understanding of the risks in historical financial research with LLMs and the need for rigorous evaluation beyond training boundaries.

4.1 Macroeconomic Data Recall

To assess GPT-4o's memorization of macroeconomic indicators, we tested its ability to recall monthly values (quarterly for GDP) across various variables, using data from January 1990 to September 2023, all within the model's training cutoff of October 2023. The indicators were divided into two groups: rates (GDP Growth, Inflation, Unemployment Rate, and 10-year Treasury Yield) and levels (Housing Starts, VIX, and Nonfarm Payrolls). For rates, we requested percentage values, evaluating accuracy through Mean Error, Mean Absolute Error,

Directional Accuracy (correctly identifying whether the rate was above a threshold value), and Directional Accuracy Change (correct direction of change from the previous period). For levels, we requested raw values, with errors calculated relative to actual levels through Mean Percent Error, Mean Absolute Percent Error, Directional Accuracy, and Directional Accuracy Change. We also examined levels over a recent 10-year period to explore potential recency effects. Performance was measured against actual values from Federal Reserve Economic Data (FRED), Yahoo Finance, and the Philly Fed Real-Time data, with results reported in Table 1 and Figure 3.

The results reveal an evident ability to recall macroeconomic data. For rates, the model demonstrates near-perfect recall, with Mean Absolute Errors ranging from 0.03% (Unemployment Rate) to 0.15% (GDP Growth) and Directional Accuracy exceeding 96% across all indicators, reaching 98% for 10-year Treasury Yield and 99% for Unemployment Rate. This result suggests that GPT-4o has memorized these percentage-based indicators with high fidelity.

For levels, the recall remains high, with Directional Accuracies between 92% and 100% for all indicators during the whole pre-training sample. Moreover, when focusing on the most recent 10-year period in the pre-training sample, performance improves dramatically–Mean Absolute Percent Errors fall to 1.06% for Housing Starts, 0.34% for VIX, and below 0.00% for Nonfarm Payrolls, with Directional Accuracy rising to 95%–100%. This recency effect indicates stronger memorization for more recent data, possibly due to denser representation in the training corpus.

The high recall accuracy for rates and recent levels underscores the memorization problem when evaluating LLMs' forecasting capabilities. The model's ability to reproduce precise macroeconomic values, especially for percentage-based indicators and recent periods, suggests that apparent forecasting success for pre-cutoff data may stem from retrieving memorized figures rather than genuine economic analysis. The weaker performance for levels over the full period, particularly for volatile indicators like Nonfarm Payrolls, hints at selective

memorization, where certain data types or time frames are less reliably retained. These findings reinforce concerns about using LLMs to analyze historical economic data, as their outputs may reflect training data exposure rather than predictive insight.

Table 1: Evaluation Metrics for Macro Indicators

This table reports a set of evaluation metrics for various macroeconomic indicators grouped into three panels: Rates, Levels, and Levels, Recent Period: Past 10 years. We ask the LLM to recall monthly values (quarterly for GDP, specific end of month date for 10-Year Treasury Yield and VIX) for each indicator. The indicators in the *Rates* panel include GDP Growth, Inflation, Unemployment Rate, and the 10-Year Treasury Yield. For these indicators, we ask the LLM to give us a percentage. The *Levels* panel includes Housing Starts, VIX, and Nonfarm Payrolls, evaluated over the full sample period. The *Levels, Recent Pre-cutoff Period: Past 10 years* panel evaluates these same indicators over a more recent, shorter period. *Mean Error (ME)*, *Mean Absolute Error (MAE)*, *Mean Percent Error (MPE)*, *Mean Absolute Percent Error (MAPE)*, *Directional Accuracy*,and *Directional Accuracy Change* are reported in percentage points (0.01 means 0.01%). For *Rates*, the *ME* is the difference $EstimatedRate - ActualRate$. *MAE* is calculated by taking the average of the absolute value of the *ME*. *Directional Accuracy* is the proportion of predictions that correctly identify whether the rate or level is above a threshold value (2.5% for GDP Growth, 3% for Inflation, 4% for Unemployment Rate, 4% for the 10-Year Treasury Yield, 16 for VIX, 1400 for Housing Starts, and 200 for Nonfarm Payrolls). For *Levels*, the *MPE* is calculated by taking the average of the percent error $(EstimatedLevel - ActualLevel)/ActualLevel$. *MAPE* is calculated by taking the average of the absolute value of the percent error. *Directional Accuracy Change* is the proportion of predictions that correctly identify the direction of change (up or down) relative to the previous month. *Confidence Calibration* is the correlation between the LLM's confidence level (on a scale from 0 to 100) and the MAPE. *Num Obs* is the number of observations used in the evaluation, *Start Date* and *End Date* indicate the period over which the metrics were computed. *Refusals* are the number of instances in which the model withheld a prediction by either answering "null" or 0.

Panel A: Rates	ME (%)	MAE (%)	Directional Accuracy (%)	Directional Accuracy Change (%)	Confidence Calibration	Start Date	End Date	Num Obs	Refusals
Pre-cutoff									
GDP Growth	0.01	0.15	96.27	96.99	-0.27	01/01/1990	07/01/2023	134	0
Inflation	0.00	0.04	98.02	93.07	-0.11	01/01/1990	09/01/2023	405	3
Unemployment Rate	-0.00	0.03	99.26	83.46	0.09	12/01/1989	09/01/2023	406	0
10-Yr Treasury Yield	-0.00	0.06	98.52	88.12	-0.40	01/31/1990	09/29/2023	405	0
Post-cutoff									
GDP Growth	-0.46	0.66	40.00	100.00	N/A	12/01/2023	12/01/2024	5	0
Inflation	0.35	0.38	47.06	56.25	0.70	10/01/2023	02/01/2025	17	0
Unemployment Rate	-0.20	0.26	52.94	31.25	0.47	10/01/2023	02/01/2025	17	0
10-Yr Treasury Yield	-0.26	0.49	27.78	47.06	0.14	10/31/2023	03/07/2025	18	0

Panel B: Levels	MPE (%)	MAPE (%)	Directional Accuracy (%)	Directional Accuracy Change (%)	Confidence Calibration	Start Date	End Date	Num Obs	Refusals
Pre-cutoff									
VIX	10.42	13.62	100.00	83.94	-0.51	01/02/1990	09/29/2023	443	16
Housing Starts	-2.38	3.93	100.00	81.86	-0.28	01/01/1990	09/01/2023	398	0
Nonfarm Payrolls	-7.67	66.47	92.80	94.03	-0.11	01/01/1990	09/01/2023	403	1
Recent Pre-cutoff Period: Past 10 years									
VIX	0.04	0.34	100.00	98.13	-0.20	10/31/2014	09/29/2023	108	0
Housing Starts	-0.22	1.06	95.28	98.10	-0.14	10/01/2014	09/01/2023	106	0
Nonfarm Payrolls	-0.00	0.00	100.00	100.00	-0.14	10/01/2014	09/01/2023	108	0
Post-cutoff									
VIX	16.87	21.14	50.00	61.54	N/A	10/31/2023	02/28/2025	14	3
Housing Starts	2.25	8.54	35.29	56.25	N/A	10/01/2023	02/01/2025	17	0
Nonfarm Payrolls	69.51	97.44	58.82	68.75	-0.21	10/01/2023	02/01/2025	17	0

4.2 Market Index Recall

We next evaluate GPT-4o's memorization of market index data by testing its ability to recall daily and monthly values for the S&P 500, Dow Jones Industrial Average (DJIA), and Nasdaq Composite, using data from December 1989 to March 2025. For numerical recall tests, we requested exact closing values at daily frequency, both without context and with the previous two days' levels provided, as well as monthly returns. Additionally, we assessed directional changes (up or down) and relative performance between index pairs at monthly frequency. Performance metrics include Mean Percent Error, Mean Absolute Percent Error, and Directional Accuracy Change (proportion of predictions correctly identifying the direction of change relative to the previous period) for numerical predictions and accuracy for directional and relative performance tasks, all compared against actual values from Yahoo Finance. Results are reported in Table 2, with pre-cutoff (before October 2023) and post-cutoff (after October 2023) periods distinguished to isolate memorization effects.

For pre-cutoff daily exact numerical levels, GPT-4o exhibits strong recall, with Mean Absolute Percent Errors of 0.61% for S&P 500, 0.47% for DJIA, and 1.80% for Nasdaq Composite, and Directional Accuracy Change ranging from 69.44% (Nasdaq) to 82.23% (DJIA). Providing context improves slightly accuracy for S&P 500 (0.58%) and Nasdaq (1.06%). Prompting directly for returns, which we test monthly, yields higher Directional Accuracy (79.36%–85.26%), reflecting robust memorization of directional trends. Other tests further confirm memorization: prompting directly for directional performance ("up" or "down") exceeds 79% accuracy across indices, and relative performance accuracy ranges from 82.86% (S&P 500 vs. Nasdaq) to 87.10% (S&P 500 vs. DJIA).

In contrast, post-cutoff performance collapses, with Mean Absolute Percent Errors ballooning to 13.03%–19.75% for exact levels and Directional Accuracy Change dropping to near-random levels (44.10%–49.81%), indicating no data leakage beyond the training cutoff.

The sharp pre-cutoff accuracy, particularly for exact levels and relative performance, highlights GPT-4o's extensive memorization of historical index data, posing challenges for

Figure 3: Recall of exact numerical values of macro indicators.

This figure shows the LLM's estimated values of some macro indicators compared to the actual values. Panels A, C, E, and G graph the actual values against the estimated values. Panels B, D, F, and H show the estimation error. Estimation error is calculated as *(Estimated - Actual)/Actual* and is shown in percentages (5 means 5%).

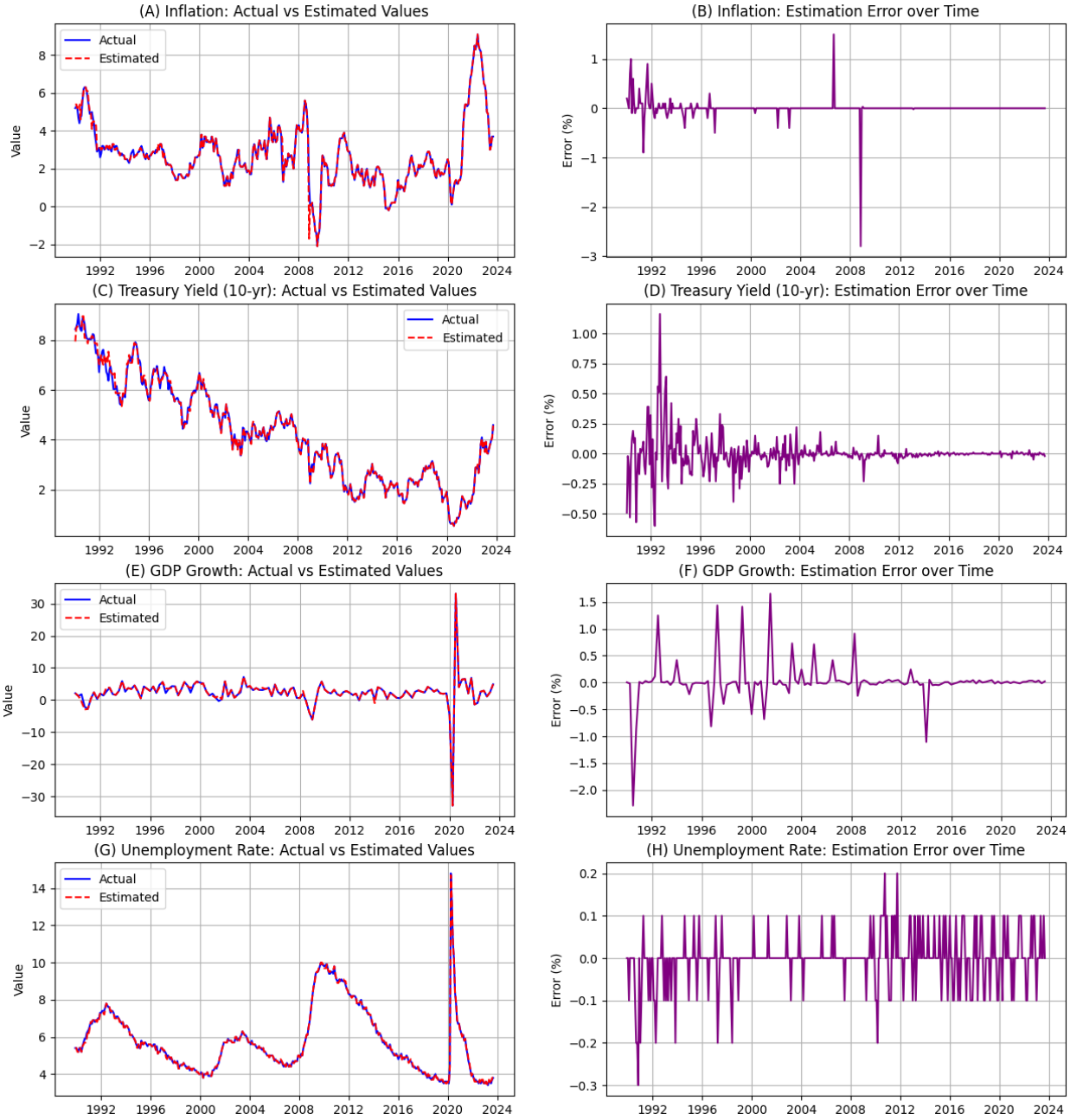


Table 2: Evaluation Metrics for Market Indices

This table reports a set of evaluation metrics assessing the LLM's ability to recall market index levels and their changes over time. These tests are done at the daily or monthly frequency. We ask the LLM to recall the closing value of the index each trading day. Panel A provides metrics for predictions of *Daily Levels* and *Daily Levels with context* (where the previous two days' index levels are provided). We ask the LLM to provide monthly returns for these indices as well. Metrics include *Mean Percent Error (MPE)*, *Mean Absolute Percent Error (MAPE)*, and *Directional Accuracy Change*, all reported in percentage points (0.10 means 0.10%). *MPE* is calculated by averaging the percent error $(EstimatedLevel - ActualLevel)/ActualLevel$. *MAPE* takes the average absolute value of the percent errors. *Directional Accuracy Change* measures the proportion of predictions correctly identifying the direction of change (up or down) relative to the previous day. *Confidence Calibration* reports the correlation between the LLM's confidence level (on a scale from 0 to 100) and mean absolute percent error. Panel B presents accuracy metrics related to predicting *Directional Changes* and *Relative Performance* between indices. *Directional Changes* asks the LLM directly for an up or down answer for each month. *Relative Performance* asks the LLM to answer which index of the index pair performed better during the month. *Accuracy* reports the proportion of predictions correctly identifying either the direction of change or relative performance in percentage points. *Confidence Calibration* in this panel reflects the correlation between the LLM's confidence and the MAPE. Results are separately provided for the S&P 500 (SP500), Dow Jones Industrial Average (DJIA), and Nasdaq Composite indices.

Panel A: Numerical Tests

	MPE (%)	MAPE (%)	Directional Accuracy Change (%)	Confidence Calibration	Num Obs	Start Date	End Date	Refusals
Daily Levels: Pre-cutoff								
SP500	0.12	0.61	80.68	-0.14	8489	12/29/1989	09/29/2023	0
DJIA	0.10	0.47	82.23	-0.35	7982	01/02/1992	09/29/2023	0
Nasdaq Composite	0.18	1.80	69.44	-0.12	8489	12/29/1989	09/29/2023	0
Daily Levels: Post-cutoff								
SP500	-16.59	16.70	45.97	-0.13	249	10/02/2023	03/05/2025	64
DJIA	-12.97	13.03	49.81	-0.15	270	10/02/2023	03/04/2025	87
Nasdaq Composite	-19.66	19.75	44.10	-0.40	230	10/02/2023	03/05/2025	62
Daily Levels: Pre-cutoff with context								
SP500	0.13	0.50	80.84	-0.16	8497	12/29/1989	09/29/2023	0
DJIA	0.05	0.43	81.60	-0.17	7983	01/06/1992	09/29/2023	510
Nasdaq Composite	0.00	1.06	68.77	-0.18	8492	12/29/1989	09/29/2023	2
Daily Levels: Post-cutoff with context								
SP500	0.02	0.63	57.09	0.03	290	10/03/2023	03/07/2025	36
DJIA	-12.97	13.03	49.81	-0.15	270	10/02/2023	03/04/2025	87
Nasdaq Composite	-19.66	19.75	44.10	-0.40	230	10/02/2023	03/05/2025	62
Monthly Returns: Pre-cutoff								
SP500	-0.70	3.35	85.26	0.27	407	12/01/1989	10/01/2023	30
DJIA	-0.70	3.25	81.10	0.29	381	02/01/1992	10/01/2023	25
Nasdaq Composite	-1.01	4.81	79.36	0.21	407	12/01/1989	10/01/2023	44

Panel B: Other Tests

	Accuracy (%)	Confidence Calibration
Monthly Directional Changes: Pre-cutoff		
SP500	82.80	0.33
DJIA	80.63	0.29
Nasdaq Composite	79.36	0.29
Monthly Relative Performance: Pre-cutoff		
SP500, DJIA	87.10	0.23
SP500, NDAQ	82.86	0.49
NDAQ, DJIA	83.87	0.42

forecasting studies. The model's ability to recall precise closing values and correctly identify directional trends within its training period suggests that any apparent predictive success may reflect memorized data rather than analytical capability. The negligible improvement from context and the complete performance drop post-cutoff reinforce that these results stem from training data exposure. These findings caution against using LLMs for historical market analysis without ensuring data is outside their training scope, as their outputs risk being artifacts of memorization rather than genuine economic foresight.

4.3 Headline Date Identification

For headline date identification, we present GPT-4o with sets of Wall Street Journal front page headlines (approximately 9 headlines per day) from our dataset of 90,123 headlines spanning December 1989 to February 2025 without revealing their publication dates. We asked the model to identify when these headlines appeared and, in a separate test variant, predict the S&P 500 level on the following trading day. Performance was evaluated using multiple accuracy metrics: year accuracy, month-and-year accuracy, exact date accuracy, mean absolute days difference, and confidence calibration. By comparing results between pre-training headlines (where memorization could occur) and post-training headlines (where memorization is impossible), we can clearly distinguish between the model's inferential abilities and its capacity to recall memorized chronological information.

We present the results in Table 3. GPT-4o demonstrates remarkable memorization of headline chronology within its training period. For pre-cutoff headlines, it achieves 98.45% accuracy in determining the correct year and 90.38% in identifying the correct month and year. Even for exact date identification, the model achieves 47.03% accuracy—significantly above chance levels. When incorrect, the model's estimates remain close to the actual date, with a mean absolute difference of 9.52 days.

In stark contrast, for headlines published after the model's training cutoff date, performance deteriorates dramatically across all metrics. Year accuracy drops to 28.81%, month-

Table 3: Evaluation Metrics for Headlines

This table reports a set of evaluation metrics assessing the LLM's ability to recall dates associated with historical headlines, along with corresponding levels of the S&P 500 index. Metrics are separated into two panels: *Headline Dates*, focusing solely on the accuracy of predicted dates, and *Headline Dates and Levels*, evaluating the accuracy of results when we prompt the LLM to give both the dates and S&P 500 levels on the next trading day. *Mean Days Difference* is the average signed difference (in days) between predicted and actual dates, while *Mean Absolute Days Difference* reports the average absolute difference. *Year Accuracy*, *Month and Year Accuracy*, and *Exact Date Accuracy* measure the percentage of predictions correctly recalling the year, the month and year, and the exact date, respectively. *Confidence Calibration* indicates the correlation between the LLM's confidence level (on a scale from 0 to 100) and the accuracy of date predictions. *Mean Percent Error S&P 500* and *Mean Absolute Percent Error S&P 500* measure the accuracy of the LLM's predicted index levels, calculated as the average and average absolute values of the percent error, respectively, and reported in percentage points (-0.01 means -0.01%). Results are provided separately for headlines from the *Pre-training Period* and the *Post-training Period*.

	Mean Days Difference	Mean Absolute Days Difference	Year Accuracy (%)	Month and Year Accuracy (%)	Exact Date Accuracy (%)	Confidence Calibration	MPE S&P 500 (%)	MAPE S&P 500 (%)
Headline Dates								
Pre-training Period	-0.77	9.52	98.45	90.38	47.03	-0.10		
Post-training Period	413.46	414.54	28.81	20.71	7.86	-0.12		
Headline Dates and Levels								
Pre-training Period	-1.30	9.63	98.50	90.31	39.31	-0.10	0.00	0.01
Post-training Period	456.84	457.13	26.20	19.47	5.53	0.29	-0.21	0.22

and-year accuracy falls to 20.71%, and exact date accuracy declines to just 7.86%. The mean absolute difference increases to 414.54 days, indicating essentially random guessing.

We observed a similar pattern when we extended our test to ask the model to provide both the headline date and the corresponding S&P 500 level on the next trading day. For the pre-training period, the model achieved high temporal accuracy while maintaining near-perfect recall of index values (mean absolute percent error of just 0.01%). For post-training headlines, both date identification and index level predictions became significantly less accurate.

The sharp performance discontinuity at the training cutoff date provides compelling evidence that the model's apparent "knowledge" of financial chronology stems primarily from memorization rather than inference or reasoning. This finding raises significant concerns about using LLMs to analyze historical relationships between news events and market movements within their training period, as their responses may reflect memorized associations rather than genuine analytical insights.

4.4 Individual Stock Price Recall

To further investigate GPT-4o's memorization capabilities, we tested its ability to recall end-of-month closing prices for the Magnificent 7 stocks (META, GOOGL, AMZN, TSLA, NVDA, MSFT, AAPL) from January 1989 to September 2023, all within the model's training cutoff of October 2023. We queried prices both without context and with the previous two months' closing prices provided, using data from the Center for Research in Security Prices (CRSP). Performance was evaluated using Mean Percent Error, Mean Absolute Percent Error, and Directional Accuracy (correctly identifying the direction of change relative to the previous month), with results compared against actual closing prices. Table 4 reports these metrics, complementing our earlier findings on market indices by examining individual security-level memorization. We also plot the actual vs estimated values in Figures 4 and 5.

The results reveal varying recall accuracy across stocks, with notable improvements when context is provided. Without context (Panel A), GPT-4o performs best for newer stocks like META, with a Mean Absolute Percent Error of 0.37% and Directional Accuracy Change of 99.26%, but struggles with older stocks like AAPL (38.21% error, 72.68% accuracy) and MSFT (26.98% error, 76.62% accuracy). Errors are also high for NVDA (23.92%) and TSLA (9.99%), suggesting selective memorization tied to stock age or data prominence. With context (Panel B), accuracy improves significantly: Mean Absolute Percent Errors drop to 0.40% for META, 0.84% for GOOGL, and 5.89% for AAPL, with Directional Accuracy Change rising to 98.52%, 95.52%, and 83.91

These findings extend our market index results, highlighting that GPT-4o's memorization is not uniform across securities and is sensitive to contextual cues and stock-specific factors. The high accuracy for META and GOOGL, especially with context, parallels the model's strong recall of recent macroeconomic levels and index values, suggesting robust memorization of prominent, frequently referenced data. Conversely, larger errors for older stocks like AAPL and MSFT, even with context, align with the weaker recall of long-horizon macroeconomic levels, pointing to the potential dilution of older data in the training cor-

Table 4: Evaluation Metrics for Magnificent 7 Stocks

This table reports a set of evaluation metrics for the Magnificent 7 stocks which includes META, GOOGL, AMZN, TSLA, NVDA, MSFT, and AAPL. We ask the LLM to recall closing prices at the end of each month. *Mean Percent Error (MPE)*, *Mean Absolute Percent Error (MAPE)*, and *Directional Accuracy Change* are reported in percentage points (0.18 means 0.18%). *MPE* is calculated by taking the average of the percent error $(PredictedPrice - ActualPrice)/ActualPrice$. *MAPE* is calculated by taking the average of the absolute value of the percent error. *Directional Accuracy Change* is the proportion of predictions that went in the correct direction (up or down) with respect to the previous month. *Confidence Calibration* is the correlation between the LLM's confidence level (on a scale of 0 to 100) and the MAPE. *Num Obs* is the number of observations used in the evaluation, *Start Date* and *End Date* indicate the period over which the metrics were computed. *Refusals* are the number of instances in which the model withheld a prediction by either answering "null" or 0. Results are provided for a prompt that contains an empty context in panel A and a prompt that provides the previous two month's closing prices as context in panel B.

Panel A: No Context

	MPE (%)	MAPE (%)	Directional Accuracy Change (%)	Confidence Calibration	Num Obs	Start Date	End Date	Refusals
META	0.18	0.37	99.26	-0.08	137	05/31/2012	09/29/2023	0
GOOGL	-1.41	1.79	93.42	-0.19	229	08/31/2004	09/29/2023	1
AMZN	-5.87	7.98	91.77	-0.12	317	05/30/1997	09/29/2023	0
TSLA	-9.21	9.99	92.45	-0.13	160	06/30/2010	09/29/2023	0
NVDA	-20.60	23.92	77.05	-0.53	293	01/29/1999	09/29/2023	4
MSFT	-26.09	26.98	76.62	-0.66	403	01/31/1989	09/29/2023	14
AAPL	-36.94	38.21	72.68	-0.57	411	01/31/1989	09/29/2023	6

Panel B: With Context

	MPE (%)	MAPE (%)	Directional Accuracy Change (%)	Confidence Calibration	Num Obs	Start Date	End Date	Refusals
META	-0.18	0.40	98.52	-0.10	136	05/31/2012	09/29/2023	1
GOOGL	-0.27	0.84	95.52	0.10	224	08/31/2004	09/29/2023	6
AMZN	-0.42	3.12	93.35	0.02	317	05/30/1997	09/29/2023	0
TSLA	-0.87	2.34	94.87	-0.17	157	06/30/2010	07/31/2023	3
NVDA	1.01	8.80	80.68	-0.17	296	01/29/1999	08/31/2023	1
MSFT	1.24	4.57	84.71	-0.21	400	01/31/1990	07/31/2023	5
AAPL	-1.37	5.89	83.91	-0.25	405	01/31/1990	09/29/2023	0

Figure 4: Recall of closing prices for META, GOOGL, AMZN, and TSLA.

This figure shows the LLM's estimated closing prices for META, GOOGL, AMZN, and TSLA compared to the actual values. Panels A, C, E, and G graph the actual values against the estimated values. Panels B, D, F, and H show the estimation error. Estimation error is calculated as *(Estimated - Actual)/Actual* and is shown in percentages (5 means 5%).

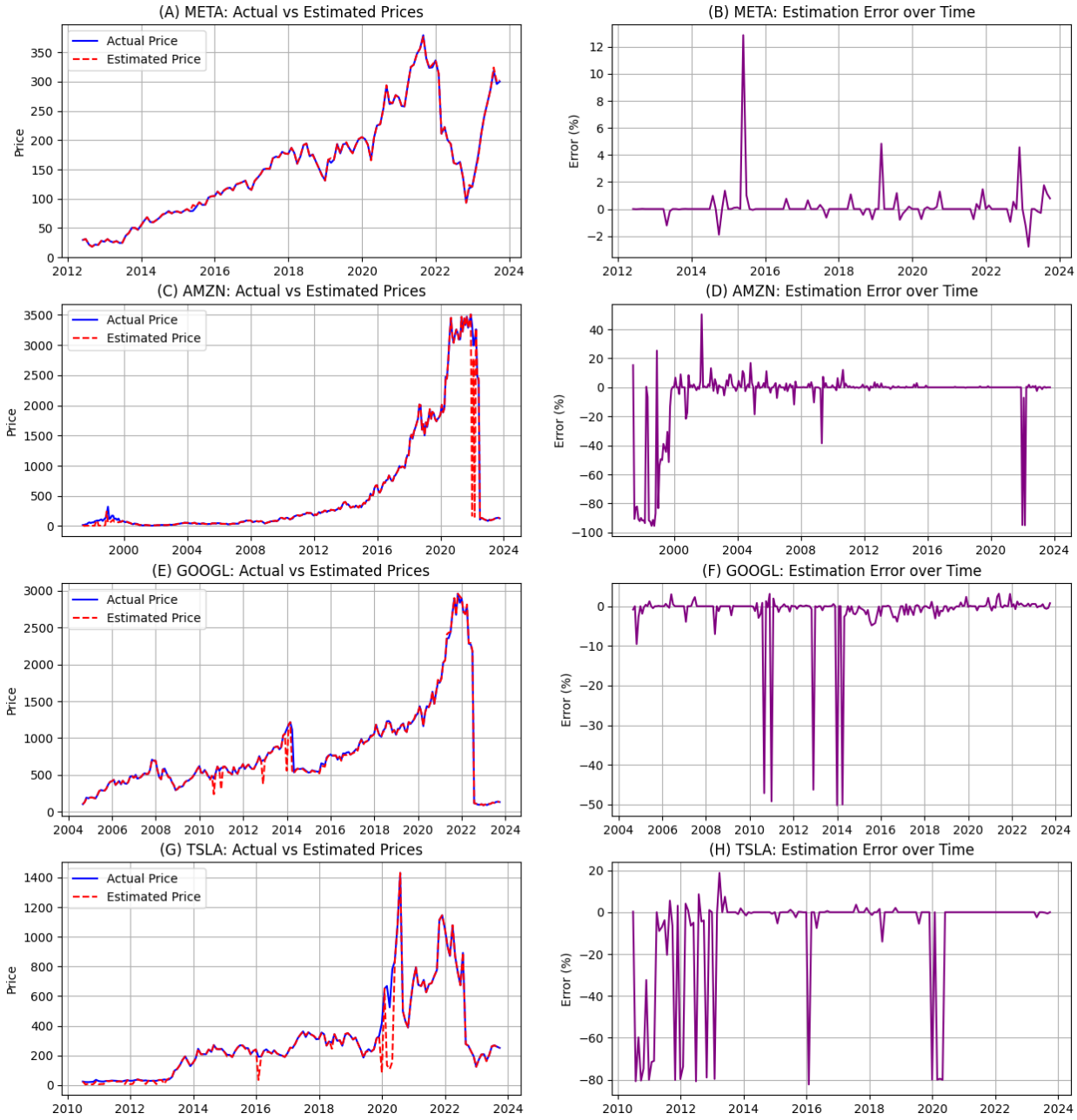


Figure 5: Recall of closing prices for NVDA, AAPL, and MSFT.

This figure shows the LLM's estimated closing prices for NVDA, AAPL, and MSFT compared to the actual values. Panels A, C, and E graph the actual values against the estimated values. Panels B, D, and F show the estimation error. Estimation error is calculated as *(Estimated - Actual)/Actual* and is shown in percentages (5 means 5%).

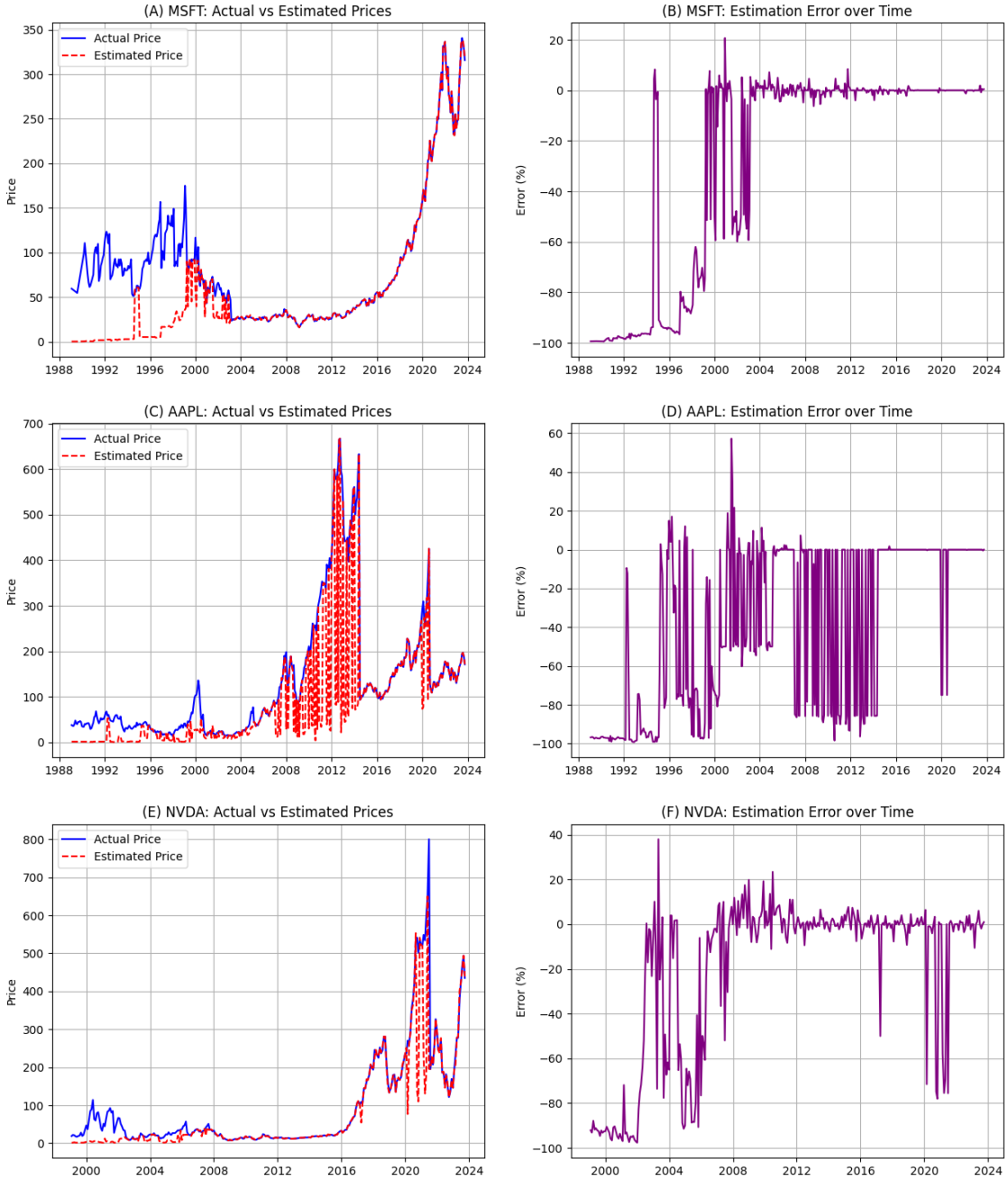
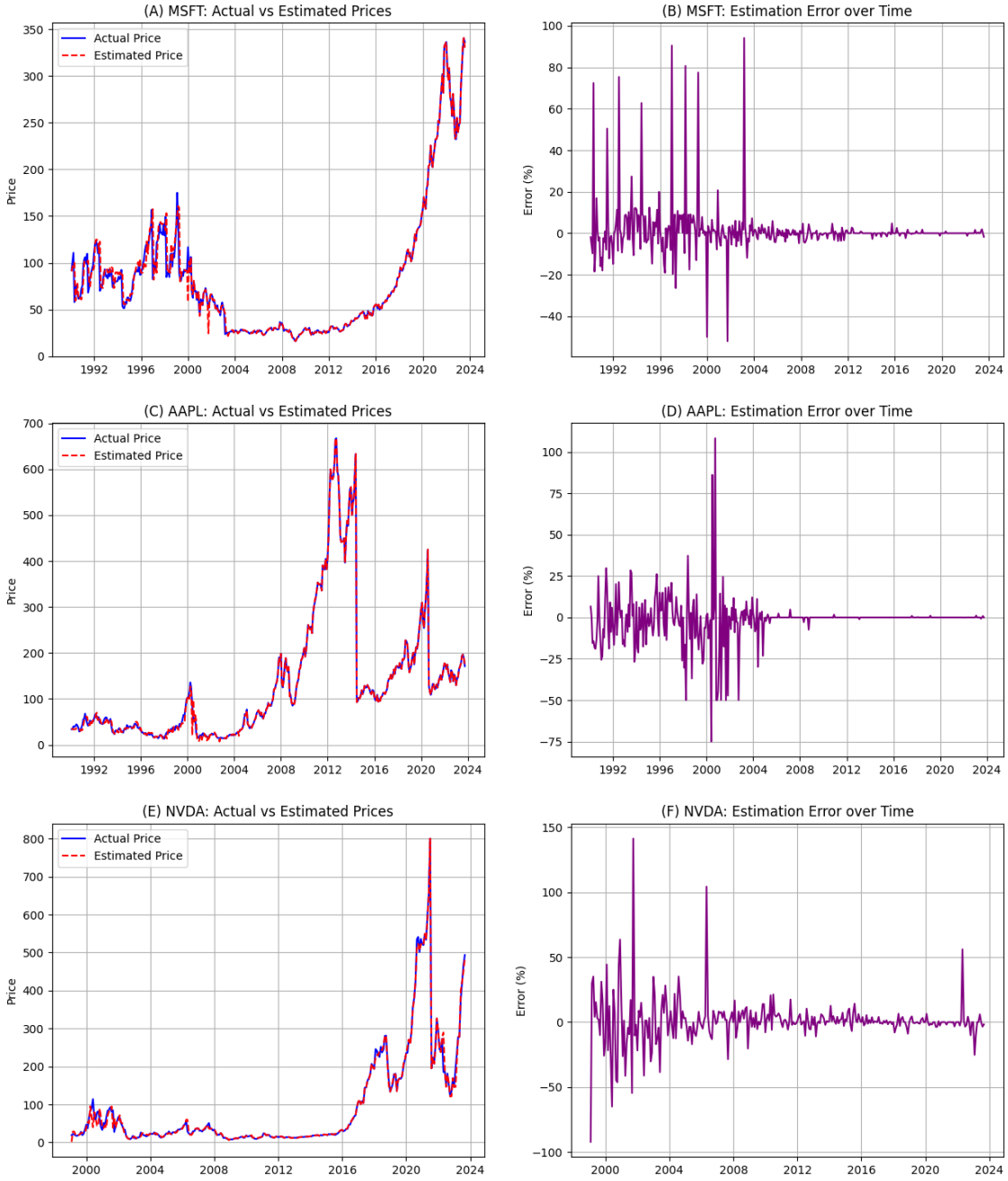


Figure 6: Recall of closing prices for NVDA, AAPL, and MSFT with context.

This figure shows the LLM's estimated closing prices for NVDA, AAPL, and MSFT compared to the actual values. We give the LLM two previous end of the month closing prices given as context. Panels A, C, and E graph the actual values against the estimated values. Panels B, D, and F show the estimation error. Estimation error is calculated as *(Estimated - Actual)/Actual* and is shown in percentages (5 means 5%).



pus. This selective memorization reinforces the challenge for financial research: apparent forecasting success for individual stocks within the training period may reflect memorized prices rather than predictive ability, necessitating evaluations with post-cutoff data to ensure methodological rigor.

Moreover, Figure 6 shows why any memorization performance only provides a lower bound. The errors substantially decrease when we give ChatGPT the prices for the previous two months. This situation is relevant as when forecasting, researchers typically provide contextual information.

4.5 Portfolio Stock Price Recall by Size

To broaden our analysis of GPT-4o's memorization, we also examine its recall of end-of-month closing prices for portfolios of stocks grouped by market capitalization—Small, Medium, and Large—covering January 1990 to September 2023, all within the model's training cutoff of October 2023. Stocks are divided into market cap terciles, with 10 stocks randomly sampled from each tercile and resampled annually to account for size changes, using stock price data from CRSP. We test recall without context, with the previous two months' prices provided, and over a recent 10-year period (2014–2023) to assess recency effects. Performance metrics, including Mean Percent Error, Mean Absolute Percent Error, and Directional Accuracy (correct direction of change relative to the prior month), are reported in Table 5. We also plot the actual vs estimated values in Figure 7.

The results show weaker recall compared to the Magnificent 7, with accuracy improving for larger stocks and recent periods. Without context, Mean Absolute Percent Errors are high: 18.35% for Small, 15.94% for Medium, and 11.36% for Large stocks, with Directional Accuracy ranging from 46.58% to 60.52%. The high refusal rates (1441 for Small, 837 for Large) suggest uncertainty for less prominent stocks. Over the recent 10-year period, errors decrease significantly—9.79% for Small, 5.94% for Medium, and 3.57% for Large—with Directional Accuracy rising to 52.54%–77.03% and near-zero refusals, echoing the recency

Table 5: Evaluation Metrics for Portfolio of Stocks by Size

This table reports a set of evaluation metrics for portfolios of stocks grouped by size: Small, Medium, and Large. We divide stocks into terciles by market cap (using the NYSE stocks only to calculate cutoffs) and sample 10 stocks from each tercile, resampling each year. We then ask the LLM to recall end of month closing prices for each stock. *Mean Percent Error (MPE)*, *Mean Absolute Percent Error (MAPE)*, and *Directional Accuracy Change* are reported in percentage points (0.78 means 0.78%). *MPE* is calculated by taking the average of the percent error $(PredictedPrice - ActualPrice)/ActualPrice$. *MAPE* is calculated by taking the average of the absolute value of the percent error. *Directional Accuracy Change* is the proportion of predictions that went in the correct direction (up or down) relative to the previous month. *Confidence Calibration* is the correlation between the LLM's confidence level (on a scale from 0 to 100) and the mean absolute percent error. *Num Obs* is the number of observations used in the evaluation. *Start Date* and *End Date* indicate the period over which the metrics were computed. *Refusals* represent the number of instances in which the model withheld a prediction by either answering "null" or 0. Results are provided for the full sample period, a recent period covering the past 10 years, and *With Context*, in which the previous two months' closing prices are provided to the model.

	MPE (%)	MAPE (%)	Directional Accuracy Change (%)	Confidence Calibration	Start Date	End Date	Num Obs	Refusals
Small	-3.50	18.35	46.58	-0.26	01/31/1991	09/29/2023	2539	1441
Medium	-2.65	15.94	51.39	-0.31	12/31/1990	09/29/2023	2859	1159
Large	-4.42	11.36	60.52	-0.37	01/31/1990	09/29/2023	3188	837
Recent Period: Past 10 years								
Small	0.32	9.79	52.54	-0.04	01/31/2014	09/29/2023	1145	4
Medium	0.77	5.94	64.84	-0.00	01/31/2014	09/29/2023	1158	3
Large	0.57	3.57	77.03	-0.07	01/31/2014	09/29/2023	1158	1
With Context								
Small	0.52	8.84	74.57	-0.03	01/31/1990	09/29/2023	3856	118
Medium	1.02	8.20	73.89	-0.04	01/31/1990	09/29/2023	3883	111
Large	1.11	6.59	74.70	-0.10	01/31/1990	09/29/2023	3803	158

effect seen in macroeconomic levels. With context, errors further improve for Directional Accuracy with all three sizes reaching at least 73%. Larger stocks consistently show better recall, likely due to greater data prominence, aligning with the Magnificent 7's stronger performance for high-profile securities.

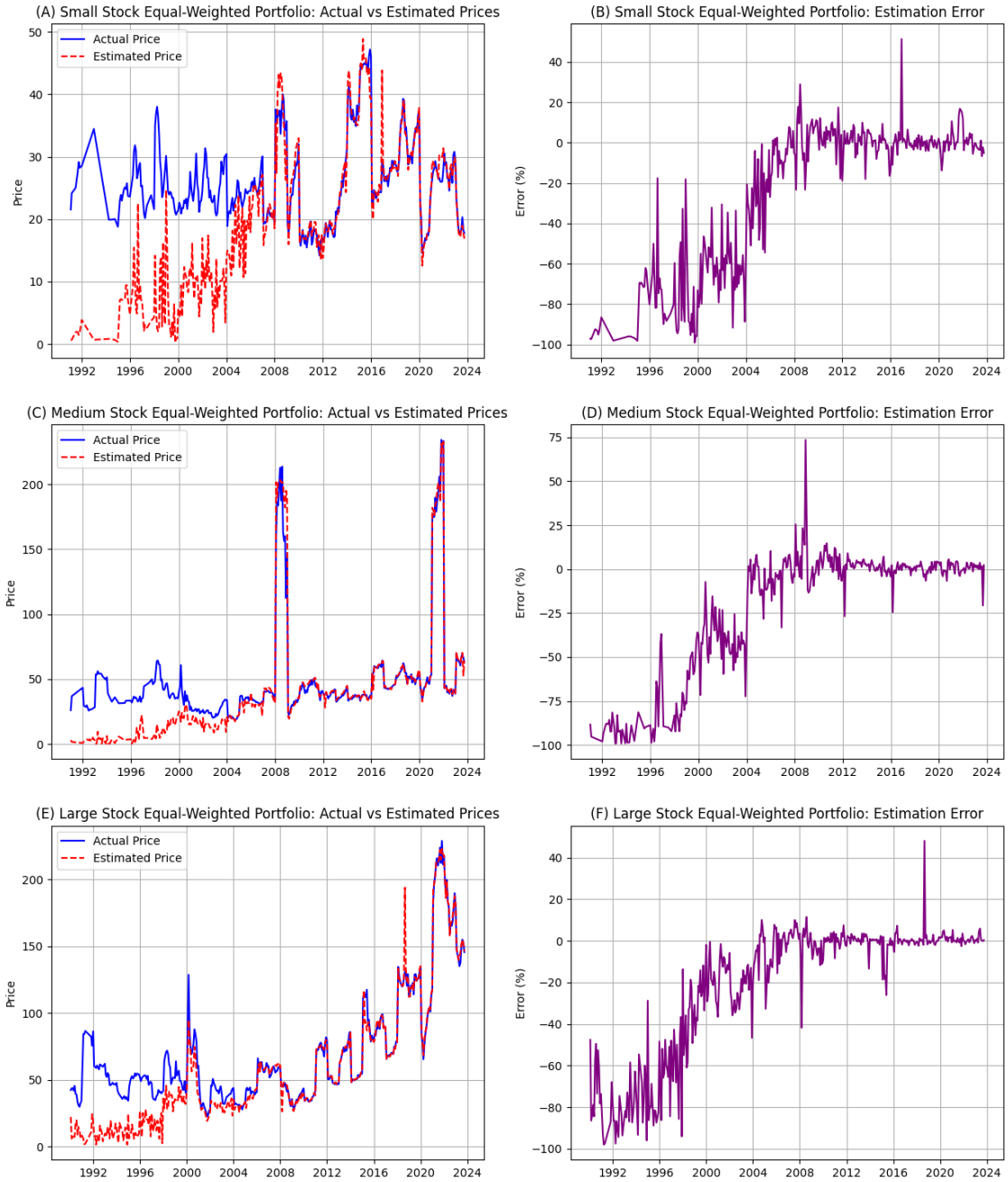
These findings complement our Magnificent 7 results, revealing that GPT-4o's memorization weakens for less prominent stocks, particularly smaller ones, and is strongest for recent, larger-cap data. The high errors and refusals for small stocks contrast with the precision for META or GOOGL, suggesting memorization is skewed toward widely covered securities, similar to the robust recall of market indices and macroeconomic rates. The recency and context effect parallel improvements seen in prior tests, but the lower directional accuracy with context indicates limits in capturing trends for diverse portfolios. This selective memorization underscores the risk of relying on LLMs for historical stock analysis, as apparent forecasting accuracy may stem from memorized prices of prominent stocks rather than broad predictive insight, reinforcing the need for post-cutoff evaluations.

4.6 Can LLMs Follow "Fake" Knowledge Cutoff Prompts?

To assess whether GPT-4o can adhere to instructions not to use information beyond a specified cutoff, we test its performance in predicting U.S. real GDP growth rates using data from the FRED database, spanning March 1990 to June 2023. We design three prompting conditions using an artificial cutoff date of December 2010: (i) one where both system and user messages explicitly restrict knowledge to pre-2010 data, (ii) one where only the system prompt imposes this fake cutoff, and (iii) one where only the user prompt imposes this fake cutoff. The task required predicting monthly year-over-year GDP growth, with data split into pre-cutoff (1990–2010) and post-cutoff (2011 onward) periods to evaluate compliance with the constraint. Performance metrics include Mean Percent Error, Mean Absolute Percent Error, Directional Accuracy (correctly identifying up/down changes), and refusal counts, reported in Table 6, building on our prior findings by directly probing the model's ability to

Figure 7: Recall of exact numerical levels of closing prices for other stocks.

This figure shows the LLM's estimated closing prices for randomly selected small, mid, and large cap stocks compared to the actual closing prices. Panels A, C, and E graph the actual values against the estimated values. Panels B, D, and F show the estimation error. Estimation error is calculated as *(Estimated - Actual)/Actual* and is shown in percentages (5 means 5%). The price plotted is the equal-weighted average price of the small, mid and large cap groups. For the large stock portfolio, ticker BRK was excluded for ease of plotting due to the extreme jump in average price with its inclusion.



isolate historical forecasting from memorized data.

When both system and user prompts enforce the pre-2010 cutoff, GPT-4o performs well on pre-cutoff data, with a Mean Absolute Percent Error of 0.27% and Directional Accuracy of 95.18%, but struggles post-cutoff, showing a higher error (0.59%) and lower accuracy (92.31%), with 38 refusals out of 51 post-cutoff observations. This high refusal rate and reduced performance suggest the model partially respects the constraint, limiting its reliance on memorized post-2010 data. We also find that the results are similar when only the user prompt specifies the 2010 cutoff. In contrast, when only the system prompt specifies the cutoff, pre-cutoff accuracy improves slightly (0.11% error, 98.80% accuracy). Still, post-cutoff performance remains implausibly strong (0.05% error, 96.00% accuracy) with only one refusal across 50 observations. This minimal drop in accuracy post-cutoff indicates that without reinforced user instructions, the model likely accesses memorized data, undermining the cutoff constraint.

These results connect to our earlier findings on macroeconomic indicators, where high pre-cutoff accuracy reflected memorization. The strong post-cutoff performance without user prompt reinforcement mirrors the suspiciously high accuracy seen in other tests when constraints were not strictly enforced, suggesting that GPT-4o defaults to using its full knowledge unless explicitly and repeatedly directed otherwise. The high refusal rate with dual prompts aligns with weaker recall for less prominent data, as seen in small-cap stocks, indicating partial compliance but not complete isolation from memorized information. This failure to fully respect cutoff instructions reinforces the challenge of using LLMs for historical forecasting, as their outputs may subtly incorporate memorized data, necessitating post-cutoff evaluations to ensure genuine predictive ability.

Table 6: Knowledge Pool Cutoff

This table reports GPT-4o's performance on U.S. real GDP growth predictions evaluated on data from the Philadelphia Fed's Real-Time Data Set. We evaluate model accuracy under different knowledge cutoff constraints: one where both system and user prompts reinforce the knowledge cutoff (pre-2010 only), another where only the system prompt specifies the cutoff, and one where only the user prompt specifies the cutoff. The task involves predicting the monthly year-over-year GDP growth rate, with test data split into pre-cutoff (1990–2010) and post-cutoff (2011 onward) periods to assess whether the model respects the stated cutoff. Metrics include Mean Error (ME), Mean Absolute Error (MAE), Directional Accuracy (percentage of guesses correctly above a threshold of 2.5%), Directional Accuracy Change (percentage of correct up/down changes), Confidence Calibration (correlation between the LLM's confidence level and the MAPE), total observations, and refusal counts. The results indicate that explicitly instructing the model not to use post-2010 data yields higher refusal rates and weaker post-cutoff performance, consistent with adherence to the knowledge constraint.

	ME (%)	MAE (%)	Directional Accuracy (%)	Directional Accuracy Change (%)	Confidence Calibration	Start Date	End Date	Num Obs	Refusals
GDP Growth: Our prompt with both system and user message knowledge cutoff									
Pre fake-cutoff	0.07	0.27	97.59	95.12	-0.02	03/01/1990	12/01/2010	83	0
Post fake-cutoff	0.58	0.59	69.23	75.00	0.51	03/01/2011	12/01/2016	13	38
GDP Growth: Our prompt with system but no user message knowledge cutoff									
Pre fake-cutoff	0.02	0.11	97.59	97.56	-0.12	03/01/1990	12/01/2010	83	0
Post fake-cutoff	-0.01	0.05	98.00	100.00	-0.24	03/01/2011	06/01/2023	50	1
GDP Growth: Our prompt with user but no system message knowledge cutoff									
Pre fake-cutoff	0.08	0.22	98.80	96.34	-0.18	03/01/1990	12/01/2010	83	0
Post fake-cutoff	0.60	0.61	66.67	70.59	-0.15	03/01/2011	06/01/2020	18	32

4.7 Testing Masking Effectiveness

To evaluate whether masking can prevent GPT-4o from accessing memorized information, we examine its ability to identify the firm, year, and quarter of anonymized earnings conference call transcripts from Capital IQ, covering July 2006 to December 2021, all within the model's training cutoff of October 2023. Transcripts were anonymized using the entity neutering approach of Engelberg et al. (2025), removing identifying details like company names and dates. We focused on the Magnificent 7 stocks (AAPL, META, MSFT, GOOG, NVDA, TSLA, AMZN) and portfolios grouped by market cap terciles (Large, Medium, Small), measuring performance through Mean Years Difference, Mean Absolute Years Difference, and accuracy in identifying the exact year, quarter and year, and firm. Results, reported in Table 7, extend our prior findings by probing whether masking mitigates the memorization seen in stock price and macroeconomic data recall.

Table 7: Anonymous Conference Calls

This table reports GPT-4o's performance in identifying the correct firm, year, and quarter of anonymized earnings conference call transcripts obtained from Capital IQ. For each firm, we report the mean and mean absolute difference between the model's predicted and actual transcript years, as well as the percentage of calls for which the model correctly identified the exact year, the exact quarter and year, and the correct firm. Results are grouped by firm and by market capitalization terciles (Large, Mid, Small). Terciles are formed using NYSE cutoffs.

Panel A: Magnificent Seven Firms

Ticker	Mean Years Difference	Mean Absolute Years Difference	Year Accuracy (%)	Quarter and Year Accuracy (%)	Firm Accuracy (%)	Num Obs
AAPL	-0.03	0.06	95.24	92.06	100.00	63
META	0.21	0.26	73.68	2.63	100.00	38
MSFT	-0.73	0.85	61.29	1.61	100.00	62
GOOG	-1.29	1.83	57.14	7.94	90.48	63
NVDA	-1.67	1.94	47.06	1.96	92.16	51
TSLA	-4.11	4.20	33.33	4.44	82.22	45
AMZN	-2.23	2.42	23.44	4.69	82.81	64

Panel B: Firms by Market Cap

Size	Mean Years Difference	Mean Absolute Years Difference	Year Accuracy (%)	Quarter and Year Accuracy (%)	Firm Accuracy (%)	Num Obs
Large	-1.75	2.12	42.05	9.09	65.53	528
Mid	-2.08	2.57	38.49	8.11	40.38	530
Small	-1.69	2.44	31.67	6.43	32.14	420

GPT-4o demonstrates a remarkable ability to deanonymize transcripts, particularly for

prominent firms. For AAPL, the model achieves 100% firm accuracy, 95.24% year accuracy, and 92.06% quarter-and-year accuracy, with a Mean Absolute Years Difference of just 0.06 years. META and MSFT also show perfect firm identification (100%), though year accuracy drops to 73.68% and 61.29%, respectively, and quarter-and-year accuracy is low (2.63% and 1.61%). Performance weakens for GOOG (90.48% firm accuracy), NVDA (92.16%), TSLA (82.22%), and AMZN (82.81%), with year accuracy ranging from 23.44% to 57.14% and Mean Absolute Years Differences rising to 1.83–4.20 years. Across market cap terciles, firm accuracy declines from 65.53% for Large to 32.14% for Small stocks, with year accuracy dropping from 42.05% for Large to 31.67% for Small stocks, indicating stronger memorization for larger, more prominent firms, consistent with our portfolio stock findings.

These results reveal that masking fails to completely prevent GPT-4o from reconstructing identifying information, paralleling the challenges in enforcing knowledge cutoffs. The high firm and year accuracy for AAPL and META aligns with their strong price recall, suggesting that prominent firms' contextual patterns—likely abundant in training data—enable deanonymization. Weaker performance for smaller stocks mirrors the higher errors for small-cap portfolios, reinforcing that memorization favors well-represented entities. The ability to identify firms and years from anonymized texts implies that LLMs can leverage subtle cues to access memorized data, undermining masking as a safeguard against forecasting contamination. This finding underscores the need for post-cutoff data to evaluate true predictive ability, as masked historical analyses may still reflect memorized outcomes rather than genuine insight.

We further evaluate whether masking can prevent GPT-4o from accessing memorized information by repeating the test for firm headlines. The firm headlines cover January 2000 to October 2023 for the pre-training period. Additionally, we have October 2023 to May 2024 in the post-training period. In contrast with the earnings call transcripts, the firm headlines are shorter texts and are more uniform in wording. After the entity neutering process, routine news such as executives trading shares of the firm or announcements of

earnings are almost indistinguishable between firms and across time.

Despite this, we show in Table 8 that this masking approach could reasonably obscure the dates of the news, but GPT-4o could guess the Magnificent 7 firms with 70% accuracy, the top decile of firms by market capitalization with 55.64% accuracy and the top quintile of firms by market capitalization with 46.71% accuracy. Even for the dates, while the LLM was not extremely accurate, it could guess the year during the pre-training period almost 20% of the time with accuracy rising to 50% near the cut-off date.

Table 8: Anonymized Firm Headlines

This table reports a set of evaluation metrics assessing the LLM's ability to recall dates and stock tickers associated with historical headlines for specific firms. The LLM was prompted with a group of anonymized headlines from RavenPack about a firm during a month and asked to state what month, year, and company the headlines were referencing. Panel A reports the results for identifying the anonymized headlines. *Mean Months Difference* is the average signed difference (in months) between predicted and actual dates, while *Mean Absolute Days Difference* reports the average absolute difference. *Year Accuracy* measures the percentage of predictions correctly recalling the year. *Firm Accuracy* measures the percentage of predictions correctly identifying the ticker associated with the headlines. Results are provided separately for firm headlines from the *Pre-training Period*, a more recent time frame from January 2014 to September 2023, the near cutoff period from January 2022 to September 2023 and the *Post-training Period*. Results are also shown by various firm sizes including the Magnificent Seven stocks, the top decile and quintile by market cap, and large stocks versus the rest categorized by NYSE market cap cutoffs. In Panel B, we provide a benchmark for date identification accuracy using the deanonymized headlines.

Panel A: Date and Firm Recall With Anonymized Headlines

	Mean Months Difference	Mean Absolute Months Difference	Year Accuracy (%)	Firm Accuracy (%)	Num Obs
Various sample periods					
Pre-training Period	37.76	55.17	19.65	21.13	8714
Pre-training Period Recent	26.34	42.44	23.44	23.21	2133
Pre-training Period Near Cutoff	-3.65	13.01	50.47	27.96	422
Post-training Period	-15.26	15.26	15.79	28.57	133
Various firm sizes					
Magnificent Seven	13.39	26.10	33.33	70.00	390
Top Decile	18.68	32.17	30.25	55.64	638
Top Quintile	23.14	37.46	25.71	46.71	1276
Large	25.99	44.36	22.59	32.68	5233
Small and Medium	43.55	62.98	15.11	8.12	1145

Panel B: Date Recall With Deanonymized Headlines

	Mean Months Difference	Mean Absolute Months Difference	Month Year Accuracy (%)	Num Obs
Pre-Training Period	1.53	3.94	76.71	8733
Pre-Training Period Recent	2.85	3.52	77.81	2181
Post-Training Period	-1.20	1.47	57.89	152

While the masking approach was more effective for firm headlines, this approach to mitigate memorization still fails to completely prevent GPT-4o from reconstructing identifying

information.

4.8 Addressing Look-Ahead Bias with Economic Logic

In previous tests, we show that the effectiveness of mitigating look-ahead bias using artificially imposed knowledge cut-off dates and anonymization through masking entities is limited. Instead, we propose a two-step process that retains the information the LLM needs to forecast while further mitigating the memorization problem. We begin with a set of firm specific headlines. For each headline, we ask the LLM to describe the effect this news may have on the firm omitting any specifics, outputting the economic logic only. Additionally, we ask the LLM to anonymize this economic logic using the masking technique. Using the set of economic logic constructed in the first step, we then ask the LLM to forecast whether the stock will be up or down giving it the anonymized economic logic.

Economic Logic Example:

Headline: Amazon Launches Kindle in Mexico

Economic Logic: The launch of Kindle in Mexico could expand Amazon's market reach, potentially increasing its revenue through access to a new customer base. This expansion may also enhance economies of scale, reducing per-unit costs and improving profitability. Additionally, entering a new market could diversify Amazon's revenue streams, mitigating risks associated with reliance on existing markets.

Anonymized Economic Logic: The launch of product_x in location_x could expand company_1's market reach, potentially increasing its revenue through access to a new customer base. This expansion may also enhance economies of scale, reducing per-unit costs and improving profitability. Additionally, entering a new market could diversify company_1's revenue streams, mitigating risks associated with reliance on existing markets.

Figure 8: An example of a headline, the economic logic generated from GPT-4o, and the anonymized economic logic masked using the entity neutering approach (Engelberg et al. 2025) using GPT-4o mini.

The results reported in 9 show the effectiveness of using the anonymized economic logic for forecasting. We tested this procedure on the Magnificent 7 stocks using GPT-4o on a

daily frequency. For each day where there was news either before 9 a.m. on the current day or after 4 p.m. the previous day, we ask the LLM to forecast whether the stock will be up or down from open to close based on the anonymized economic logic. The forecast accuracies for all seven stocks are above 50%, ranging from 50.6% to 57.9%. We also show that equal-weighted long-short strategy of a portfolio consisting of these seven stocks has an average daily return of 51 bps and an annualized sharpe ratio of 2.09.

To check the extent of the effectiveness of the anonymized economic logic in mitigating memorization, we ask GPT-4o to deanonymize the anonymized economic logic. We ask it to identify the firm and the date of the news. The results show that while GPT-4o is able to sometimes guess the identity of the firm, the date of the publications are hidden remarkably well with the LLM never being able to guess the exact date and only guessing the year between 3-9% of the time depending on the stock. Even for guessing the firm, some stocks like GOOG and NVDA were only identified accuractly 4.1% and 2.4% of the time, respectively.

Table 9: Anonymized Economic Logic

This table reports GPT-4o's performance in using the underlying economic logic of headlines to directionally forecast stock price movement on a daily frequency from January 2000 to May 2024. We use RavenPack news headlines for the Magnificent Seven stocks. For each headline, we ask GPT-4o to describe how the firm will be impacted by the news using economic logic only without specifics. We then anonymize the economic logic as another layer of abstraction away from the original text. Panel A reports the percentage accuracies of the forecasts for each firm reported in percentage points (-0.01 means -0.01%). Panel B reports the success rate of identifying the correct firm, year, and the exact date of the anonymized economic logic. Panel C reports several statistics of the different trading strategies, including the annualized Sharpe ratio, mean daily returns, standard deviation of daily returns, and maximum drawdown. The strategies include the long, short, and long-short strategy based on GPT-4o's forecasts and an equal-weight portfolio in all stocks with news the day before (regardless of news direction).

Panel A: Forecast Accuracy using Anonymized Economic Logic

Ticker	Forecast Accuracy (%)	Num Obs	Num Headlines	Avg Num Headlines Per News Day
AAPL	56.4	786	2595	3.3
META	57.9	546	872	1.6
MSFT	50.6	1216	3536	2.9
GOOG	53.7	1460	2014	1.4
NVDA	52.5	354	1041	2.9
TSLA	54.3	293	879	3
AMZN	51.1	601	1954	3.3

Panel B: Identification of Anonymized Economic Logic

Ticker	Firm Accuracy (%)	Year Accuracy (%)	Date Accuracy (%)
AAPL	45.4	4.7	0.0
META	22.0	6.3	0.0
MSFT	27.9	3.8	0.0
GOOG	4.1	5.3	0.0
NVDA	2.4	6.6	0.0
TSLA	51.8	9.2	0.0
AMZN	21.1	5.7	0.0

Panel C: Descriptive Statistics of Various Trading Strategies

	EW LS GPT 4o	EW Long Only	EW Short Only	EW All News Days
Ann. Sharpe Ratio	2.09	2.01	0.75	0.94
Daily Mean (%)	0.51	0.38	0.13	0.21
Daily Std. Dev. (%)	3.77	2.89	2.52	3.31
Max Drawdown (%)	-53.64	-46.79	-42.72	-63.33
Median number of stocks traded per day	1	1	0	1
Max number of stocks traded per day	7	6	5	7
Total number of trade days	3052	2151	1420	3052

These results show that GPT-4o was able to extract the economic logic to mitigate memorization and maintain usefulness of the information for effective forecasting.

5 Conclusion

Large language models exhibit significant memorization of economic and financial data, posing a fundamental challenge to their use in forecasting historical periods within their training

data. Through systematic testing, we document LLMs' ability to perfectly recall exact numerical values—such as S&P 500 levels, unemployment rates, and GDP figures—with high accuracy for pre-cutoff data, alongside near-perfect identification of headline dates and robust reconstruction of masked entities. This selective yet pervasive memorization can undermine the validity of LLMs' apparent forecasting accuracy, as their outputs for pre-cutoff periods are often indistinguishable from recall rather than genuine prediction.

Efforts to mitigate memorization, such as imposing artificial temporal boundaries or anonymizing data, prove inadequate. Even when explicitly instructed to ignore post-cutoff information, LLMs produce implausibly accurate forecasts for pre-cutoff periods, suggesting they bypass constraints through motivated reasoning anchored to memorized outcomes. Similarly, masking techniques fail to completely prevent LLMs from reconstructing identifying information, as they leverage subtle contextual clues to deanonymize entities like firms or periods with high success rates. These findings indicate that neither prompting strategies nor data anonymization can reliably isolate LLMs' forecasting abilities from their memorized knowledge, rendering such approaches insufficient for rigorous financial research.

To ensure methodological integrity, evaluations of LLMs' forecasting capabilities should focus exclusively on data beyond their training cutoff, where memorization is impossible. Only by testing predictions for post-cutoff periods can researchers and practitioners confidently distinguish genuine economic insight from the retrieval of memorized information. This constraint necessitates a shift in research design, prioritizing temporally consistent models or post-training data to assess LLMs' true potential in financial applications.

Applying our methodology to test for memorization in specific data used for research can give a useful lower bound on the memorization problem in a particular application. Our results underscore the necessity of reevaluating current practices in LLM-based financial research and highlight the need for robust frameworks to address the memorization problem, ensuring that claims of predictive power are grounded in actual forecasting ability rather than artifacts of training data exposure.

References

Bai, John Jianqiu, Nicole M Boyson, Yi Cao, Miao Liu, and Chi Wan. 2023. "Executives vs. chatbots: Unmasking insights through human-AI differences in earnings conference Q&A." *Northeastern U. D'Amore-McKim School of Business Research Paper,* no. 4480056.

Beckmann, Lars, Heiner Beckmeyer, Ilias Filippou, Stefan Menze, and Guofu Zhou. 2024. "Unusual Financial Communication: ChatGPT, Earnings Calls, and Financial Markets." Pre-published, January 15, 2024. SSRN Scholarly Paper. Accessed April 15, 2025. https://doi.org/10.2139/ssrn.4699231. Social Science Research Network: 4699231. https://papers.ssrn.com/abstract=4699231.

Bond, Shaun A, Hayden Klok, and Min Zhu. 2024. "Large language models and financial market sentiment." *Available at SSRN 4584928.*

Breitung, Christian, and Sebastian Müller. 2025. "Global Business Networks." *Journal of Financial Economics* 166 (April 1, 2025): 104007.

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

Cao, Yi, Long Chen, Jennifer Tucker, and Chi Wan. 2025. "Can Generative AI Help Identify Peer Firms?" Pre-published, April 4, 2025. SSRN Scholarly Paper. Accessed April 15, 2025. https://doi.org/10.2139/ssrn.4761624. Social Science Research Network: 4761624. https://papers.ssrn.com/abstract=4761624.

Chen, Jian, Guohao Tang, Guofu Zhou, and Wu Zhu. 2023. "ChatGPT and Deepseek: Can They Predict the Stock Market and Macroeconomy?" Pre-published, July 31, 2023. SSRN Scholarly Paper. Accessed April 11, 2025. https://doi.org/10.2139/ssrn.4660148. Social Science Research Network: 4660148. https://papers.ssrn.com/abstract=4660148.

Chen, Shuaiyu, T. Clifton Green, Huseyin Gulen, and Dexin Zhou. 2024. "What Does ChatGPT Make of Historical Stock Returns? Extrapolation and Miscalibration in LLM Stock Return Forecasts." Pre-published, August 30, 2024. SSRN Scholarly Paper. Accessed April 7, 2025. https://doi.org/10.2139/ssrn.4941906. Social Science Research Network: 4941906. https://papers.ssrn.com/abstract=4941906.

Chen, Yifei, Bryan T. Kelly, and Dacheng Xiu. 2022. "Expected Returns and Large Language Models." Pre-published, November 22, 2022. SSRN Scholarly Paper. Accessed January 26, 2025. Social Science Research Network: 4416687. https://papers.ssrn.com/abstract=4416687.

Degen, Dominik, Dr Jens Kengelbach, Daniel Kim, Soenke Sievers, and Yiran Wang. 2024. "Large Language Models and M&A: Can ChatGPT Help Forecast M&A Activity?" Pre-published, June 27, 2024. SSRN Scholarly Paper. Accessed April 11, 2025. https://doi.org/10.2139/ssrn.4862121. Social Science Research Network: 4862121. https://papers.ssrn.com/abstract=4862121.

Engelberg, Joseph, Asaf Manela, William Mullins, and Luka Vulicevic. 2025. "Entity Neutering." Pre-published, March 17, 2025. SSRN Scholarly Paper. Accessed April 7, 2025. Social Science Research Network: 5182756. https://papers.ssrn.com/abstract=5182756.

Glasserman, Paul, and Caden Lin. 2023. "Assessing Look-Ahead Bias in Stock Return Predictions Generated By GPT Sentiment Analysis." Pre-published, September 29, 2023. Accessed April 7, 2025. https://doi.org/10.48550/arXiv.2309.17322. arXiv: 2309.17322 [q-fin]. http://arxiv.org/abs/2309.17322.

Hansen, Anne Lundgaard, John J. Horton, Sophia Kazinnik, Daniela Puzzello, and Ali Zarifhonarvar. 2024. "Simulating the Survey of Professional Forecasters." Pre-published, December 1, 2024. SSRN Scholarly Paper. Accessed April 14, 2025. https://doi.org/10.2139/ssrn.5066286. Social Science Research Network: 5066286. https://papers.ssrn.com/abstract=5066286.

He, Songrun, Linying Lv, Asaf Manela, and Jimmy Wu. 2025. "Chronologically Consistent Large Language Models." Pre-published, March 18, 2025. Accessed April 7, 2025. https://doi.org/10.48550/arXiv.2502.21206. arXiv: 2502.21206 [q-fin]. http://arxiv.org/abs/2502.21206.

Horton, John J. 2023. "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" *SSRN Electronic Journal.*

Jha, Manish, Jialin Qian, Michael Weber, and Baozhong Yang. 2024. "Harnessing Generative AI for Economic Insights." Pre-published, October 4, 2024. SSRN Scholarly Paper. Accessed April 15, 2025. https://doi.org/10.2139/ssrn.4976759. Social Science Research Network: 4976759. https://papers.ssrn.com/abstract=4976759.

———. 2025. "ChatGPT and Corporate Policies." Pre-published, February 28, 2025. SSRN Scholarly Paper. Accessed April 11, 2025. https://doi.org/10.2139/ssrn.4521096. Social Science Research Network: 4521096. https://papers.ssrn.com/abstract=4521096.

Levy, Bradford. 2024. "Caution Ahead: Numerical Reasoning and Look-ahead Bias in AI Models." Pre-published, December 25, 2024. SSRN Scholarly Paper. Accessed April 7, 2025. https://doi.org/10.2139/ssrn.5082861. Social Science Research Network: 5082861. https://papers.ssrn.com/abstract=5082861.

Lopez-Lira, Alejandro, and Yuehua Tang. 2023. "Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models." *SSRN Electronic Journal* (April 6, 2023).

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan. 2025. "Large Language Models: An Applied Econometric Framework." Pre-published, January. Working Paper. Accessed April 7, 2025. https://doi.org/10.3386/w33344. National Bureau of Economic Research: 33344. https://www.nber.org/papers/w33344.

Manning, Benjamin S., Kehang Zhu, and John J. Horton. 2024. "Automated Social Science: Language Models as Scientist and Subjects." (Cambridge, MA) (April 29, 2024).

Pham, Van H., and Scott Cunningham. 2024. "Can Base ChatGPT Be Used for Forecasting without Additional Optimization?" Pre-published, July 3, 2024. SSRN Scholarly Paper. Accessed April 11, 2025. https://doi.org/10.2139/ssrn.4792918. Social Science Research Network: 4792918. https://papers.ssrn.com/abstract=4792918.

Rahimikia, Eghbal, and Felix Drinkall. 2024. "Re(Visiting) Large Language Models in Finance." Pre-published, September 21, 2024. SSRN Scholarly Paper. Accessed April 20, 2025. https://doi.org/10.2139/ssrn.4963618. Social Science Research Network: 4963618. https://papers.ssrn.com/abstract=4963618.

Ross, Jillian, Yoon Kim, and Andrew W. Lo. 2024. "LLM Economicus? Mapping the Behavioral Biases of LLMs via Utility Theory." Pre-published, August 5, 2024. Accessed April 7, 2025. https://doi.org/10.48550/arXiv.2408.02784. arXiv: 2408.02784 [cs]. http://arxiv.org/abs/2408.02784.

Sarkar, Suproteem, and Keyon Vafa. 2024. "Lookahead Bias in Pretrained Language Models" [in en]. *SSRN Electronic Journal.*

Sarkar, Suproteem K. 2024. "StoriesLM: A Family of Language Models With Time-Indexed Training Data." Pre-published, March 10, 2024. SSRN Scholarly Paper. Accessed April 7, 2025. Social Science Research Network: 4881024. https://papers.ssrn.com/abstract=4881024.

Tan, Lin, Huihang Wu, and Xiaoyan Zhang. 2024. "Large Language Models and Return Prediction in China." Pre-published, November 7, 2024. SSRN Scholarly Paper. Accessed April 15, 2025. https://doi.org/10.2139/ssrn.4712248. Social Science Research Network: 4712248. https://papers.ssrn.com/abstract=4712248.

Van Binsbergen, Jules H., Xiao Han, and Alejandro Lopez-Lira. 2022. "Textual Analysis of Short-seller Research Reports." Pre-published, December 22, 2022. SSRN Scholarly Paper. Accessed April 15, 2025. https://doi.org/10.2139/ssrn.3965873. Social Science Research Network: 3965873. https://papers.ssrn.com/abstract=3965873.

A Appendix

Table A1: Summary Statistics

This table reports summary statistics for stock indices including the S%P 500 (SP500), the Dow Jones Industrial Average (DJIA), and NASDAQ Composite in Panel A. We report the mean and standard deviation of the daily and monthly returns from January 1990 to February 2025. We also report the Directional Change (whether the price went up or down from one period to the next). In Panel B, we report these same statistics for the monthly returns from January 1990 to December 2023 of the Magnificent 7 stocks (GOOGL, AMZN, AAPL, MSFT, META, NVDA, TSLA) and the equal-weighted portfolios of the randomly drawn small, mid, and large stocks. In Panel C, we report the mean, standard deviation, Direction (whether the rate was higher or lower than a specified threshold), Directional Change for GDP Growth, Inflation, Unemployment Rate, and 10-Year Treasury Yield. The thresholds we use for Direction are 2.5%, 3%, 4%, and 4%, respectively. We also include the Mean Error and Mean Absolute Error when using the average as the estimate over the period of January 1990 to February 2025. In Panel D, we report the mean, standard deviation, Direction, and Directional Change for the VIX, Housing Starts, and Change in Nonfarm Payrolls from January 1990 to February 2025. The thresholds for Direction are 16, 1400, and 200, respectively. We also include the Mean Percent Error and Mean Absolute Percent Error when using the average as the estimate.

Panel A: Stock Indices	Mean of Return (%)	SD of Return (%)	Directional Change (%)
SP500 Daily	0.04	0.11	53.59
DJIA Daily	0.05	0.15	54.97
NASDAQ Composite Daily	0.05	0.15	54.97
SP500 Monthly	0.78	0.43	64.13
DJIA Monthly	1.10	0.62	61.52
NASDAQ Composite Monthly	1.10	0.62	61.52

Panel B: Stocks	Mean of Return (%)	SD of Return (%)	Directional Change (%)
GOOGL Monthly	1.46	1.15	60.09
AMZN Monthly	2.53	1.67	58.23
AAPL Monthly	1.57	1.36	55.85
MSFT Monthly	1.07	1.10	58.96
META Monthly	2.34	1.12	61.76
NVDA Monthly	3.19	2.03	58.22
TSLA Monthly	3.45	1.94	52.20
Small Stocks Monthly	1.56	1.59	53.33
Mid Stocks Monthly	0.49	1.34	51.70
Large Stocks Monthly	0.43	1.27	52.96

Panel C: Macro Rates	Mean (%)	SD (%)	Direction (%)	Directional Change (%)	Mean Error (%)	Mean Absolute Error (%)
GDP Growth	2.39	4.48	50.36	44.53	-0.00	1.93
Inflation	2.70	1.62	31.28	42.52	-0.00	1.13
Unemployment Rate	5.71	1.75	84.16	31.99	0.00	1.34
10-Year Treasury Yield	4.23	1.97	52.01	48.82	0.00	1.62

Panel D: Macro Levels	Mean	SD	Direction (%)	Directional Change (%)	Mean Percent Error (%)	Mean Absolute Percent Error (%)
VIX	19.86	7.15	65.62	47.48	11.25	29.07
Housing Starts	1323.95	384.55	45.32	50.36	12.13	29.85
Change in Nonfarm Payrolls	104.16	1062.12	37.05	43.57	-58.81	188.41

B Appendix: Prompts

B.1 Indices

Context

Context: The closing price of {data_name} was {previous_closing_price_1} on {previous_period_1} and {previous_closing_price_2} on {previous_period_2}.

Prompt: Daily Value Estimation

{context}

What was the {data_name} closing value on {date}? Provide your best estimate.

Provide a precise numerical answer. Indicate your level of confidence. Format as a JSON object with the following fields:

- **answer**: The precise numerical answer to the question. No strings.
- **confidence**: A number between 0 and 100 indicating your confidence in the answer.

Prompt: Monthly Directional Prediction

{context}

Was the {data_name} up or down for the month of {month}, {year}? Provide your best estimate.

Provide an answer that is either "up" or "down". Indicate your level of confidence. Format as a JSON object with the following fields:

- **answer**: An answer to the question that is either "up" or "down".

- **confidence**: A number between 0 and 100 indicating your confidence in the answer.

Prompt: Monthly Percentage Change

{context}

By what percentage did the {data_name} change for the month of {month}, {year}? Provide your best estimate.

Provide a precise numerical answer. Indicate your level of confidence. Format as a JSON object with the following fields:

- **answer**: The precise numerical answer to the question. No strings.
- **confidence**: A number between 0 and 100 indicating your confidence in the answer.

Prompt: Relative Performance Comparison

{context}

Which performed better in {year}: {data_name} or {data_name2}? Provide your best estimate.

Provide an answer that is either {data_name} or {data_name2}. Indicate your level of confidence. Format as a JSON object with the following fields:

- **answer**: An answer to the question that is either {data_name} or {data_name2}.
- **confidence**: A number between 0 and 100 indicating your confidence in the answer.

B.2 Headlines

Prompt: News-Based Context Construction

The model was given the following context generated from Wall Street Journal headlines published on the same day:

> Here are headlines from the Wall Street Journal written on the same day:
> {title_1}: {content_1}
> {title_2}: {content_2}
> ...

Prompt Type: Date Inference

{context}

What is the date of these headlines? Provide your best estimate.

Provide a precise date. Indicate your level of confidence. Format as a JSON object with the following fields:

- **answer**: The precise date in the format "mm/dd/yyyy".
- **confidence**: A number between 0 and 100 indicating your confidence in the answer.

Prompt Type: Market Level Prediction

{context}

First, infer the date of these headlines. What was the closing value of the {data_name} for the next trading day? Provide your best estimate.

You must provide a precise numerical answer. Indicate your level of confidence. Format as a JSON object with the following fields:

- **date**: The date of the headlines in the format "mm/dd/yyyy".

- **answer**: The precise numerical answer to the question. No strings.
- **confidence**: A number between 0 and 100 indicating your confidence in the answer.

B.3 Macro

Prompt: Monthly Rate

What was the {`data_name`} in {`month`}, {`year`}? Provide your best estimate.

Provide a precise numerical answer in percentage format. Indicate your level of confidence. Format as a JSON object with the following fields:

- **answer**: The precise numerical answer in percentage format to the question. No strings.
- **confidence**: A number between 0 and 100 indicating your confidence in the answer.

Prompt: Monthly Level

What was the {`data_name`} in {`month`}, {`year`}? Provide your best estimate.

Provide a precise numerical answer. Indicate your level of confidence. Format as a JSON object with the following fields:

- **answer**: The precise numerical answer to the question. No strings.
- **confidence**: A number between 0 and 100 indicating your confidence in the answer.

Prompt: Quarterly Rate

What was the {`data_name`} in {`quarter`} {`year`}? Provide your best estimate.

Provide a precise numerical answer in percentage format. Indicate your level of confidence. Format as a JSON object with the following fields:

- **answer**: The precise numerical answer in percentage format to the question. No strings.
- **confidence**: A number between 0 and 100 indicating your confidence in the answer.

Prompt: End-of-Month Level

What was the {data_name} on {end_of_month_date}? Provide your best estimate.

Provide a precise numerical answer. Indicate your level of confidence. Format as a JSON object with the following fields:

- **answer**: The precise numerical answer to the question. No strings.
- **confidence**: A number between 0 and 100 indicating your confidence in the answer.

B.4 Individual Stocks

Context

Context: The closing price of {data_name} was {previous_closing_price_1} on {previous_period_1} and {previous_closing_price_2} on {previous_period_2}.

(If only one previous closing price exists: The closing price of {data_name} was {previous_closing_price_1} on {previous_period_1}.)

Prompt: End-of-Month Closing Price

{context}

What was the closing price of {ticker_str} on {end_of_month_date}? Provide your best estimate.

Provide a precise numerical answer. Indicate your level of confidence. Format as a JSON object with the following fields:

- **answer**: The precise numerical answer to the question. No strings.
- **confidence**: A number between 0 and 100 indicating your confidence in the answer.

B.5 Knowledge Cutoff

Prompt: Both System and User Cutoff

System message: You're a helpful assistant who always tries to answer to the best of your recollection of facts. Knowledge cutoff: December 31st, 2010 Current date: October 1st, 2023.

User message: For the following tasks, please restrict your knowledge pool to the end of 2010. What was the earliest estimate of the US GDP growth rate in quarter? Provide your best estimate.

Provide a precise numerical answer in percentage format. Indicate your level of confidence. Format as a JSON object with the following fields:

- **answer**: The precise numerical answer in percentage format to the question. No strings.
- **confidence**: A number between 0 and 100 indicating your confidence in the answer.

Prompt: Only System Cutoff

System message: You're a helpful assistant who always tries to answer to the best of your recollection of facts. Knowledge cutoff: December 31st, 2010 Current date: October 1st, 2023.

User message: What was the earliest estimate of the US GDP growth rate in `quarter`? Provide your best estimate.

Provide a precise numerical answer in percentage format. Indicate your level of confidence. Format as a JSON object with the following fields:

- **answer**: The precise numerical answer in percentage format to the question. No strings.
- **confidence**: A number between 0 and 100 indicating your confidence in the answer.

Prompt: Only User Cutoff

System message: You're a helpful assistant who always tries to answer to the best of your recollection of facts.

User message: For the following tasks, please restrict your knowledge pool to the end of 2010. What was the earliest estimate of the US GDP growth rate in `quarter`? Provide your best estimate.

Provide a precise numerical answer in percentage format. Indicate your level of confidence. Format as a JSON object with the following fields:

- **answer**: The precise numerical answer in percentage format to the question. No strings.
- **confidence**: A number between 0 and 100 indicating your confidence in the answer.

B.6 Anonymized Conference Calls

Prompt: Anonymization, adapted from Engelberg et al. (2025)

Your role is to **ANONYMIZE** all text that is provided by the user. After you have anonymized a text, **NOBODY**, not even an expert financial analyst, should be able to read the text and know the identity of the company nor the industry the company operates in.

For example, if the text is: *The country's largest phone producer Apple had great phone related earnings but Google did not in 2024 likely because of Apple's slogan Think Different*, then you should ANONYMIZE it to:

The country's largest `product_type_1` *producer* `Company_1` *had great* `product_type_1` *related earnings but* `Company_2` *did not in* `time_1` *likely because of* `Company_1`*'s slogan* `slogan_1`.

You should also ANONYMIZE any other information which one could use to identify the company or make an educated guess at its identity. Stock tickers are identifiers and are usually four capitalized letters or less (consider TIK as a stand-in for an arbitrary ticker) and are sometimes referenced in the text in the following formats: `SYMBOL:TIK`, `TIK`, `>TIK`, `$TIK`, `$ TIK`, `SYMBOL TIK`, `SYMBOL: TIK`, `$> TIK`.

Make sure you censor `TIK` to `ticker_x`, and any other identifiers related to companies. This includes the names of individuals, locations, industries, sectors, product names and types, generic product lines, services, times, years, dates, and all numbers and percentages in the text including units. These should be replaced with: `name_x`, `location_x`, `industry_x`, `sector_x`, `product_x`, `product_type_x`, `product_line_x`, `service_x`, `time_x`, `year_x`, `date_x`, and `number_a`, `number_b`, `number_c`, respectively.

Also replace any website or internet links with `link_x`. Anonymize all location references, including cities, countries, regions, and other geographical indicators, as `location_x`. Re-

place all references to specific industries, sectors, and markets with `industry_x`, `sector_x`, or `market_x`, respectively. Replace all references to dates, times, years, quarters, months, or any other temporal markers with `date_x`, `time_x`, `year_x`, or `quarter_x`.

Replace all numeric references, including numbers, percentages, financial figures, units of measurement, ratios, revenues, margins, forecasts, and any other numeric value with anonymized markers (e.g., `number_a`, `number_b`, `number_c`). Replace all domain names and URLs with `link_x` (e.g., "ToysRUs.com" to "`link_x`"). Replace all references to specific services, stores, or platforms with `service_x` (e.g., "Amazon Prime" to "`service_x`").

You should never just delete an identifier; instead, always replace it with an anonymous analog. After you read and ANONYMIZE the text, you should output the anonymized text and nothing else.

`[Opening Statement]`

Prompt: Identification, adapted from Engelberg et al. (2025)

You will receive a body of text which has been anonymized. You are omniscient. Use all your knowledge and the context to identify which company and industry the text is about, as well as the quarter and year it was written. Make your best guess based on information and context if you are unsure. Please only provide the ticker of the company you have identified. Provide your estimate exactly in the following format, with no other text at all (TIK is your estimate of the ticker, Industry Name is your estimate of the industry, Q is your estimate of the quarter, Y is your estimate of the year): Company Estimate: TIK, Industry Estimate: Industry Name, Quarter Estimate: Q, Year Estimate: Y

`[Anonymized Opening Statement]`