A Test of Lookahead Bias in LLM Forecasts

Zhenyu Gao, Wenxi Jiang, Yutong Yan *

December 2025

Abstract

We develop a statistical test to detect lookahead bias in economic forecasts generated by large language models (LLMs). Using state-of-the-art pre-training data detection techniques, we estimate the likelihood that a given prompt appeared in an LLM's training corpus, a statistic we term Lookahead Propensity (LAP). We formally show that a positive correlation between LAP and forecast accuracy indicates the presence and magnitude of lookahead bias, and apply the test to two forecasting tasks: news headlines predicting stock returns and earnings call transcripts predicting capital expenditures. Our test provides a cost-efficient, diagnostic tool for assessing the validity and reliability of LLM-generated forecasts.

*Gao, Jiang, and Yan are at the Department of Finance, CUHK Business School, The Chinese University of Hong Kong. For helpful comments, we thank Chengwang Liao, Ron Kaniel, and seminar participants at CUHK. Our correspondences are gaozhenyu@baf.cuhk.edu.hk, wenxijiang@baf.cuhk.edu.hk, and yutong.yan@link.cuhk.edu.hk, respectively. First draft: December 2025.

1 Introduction

Large language models (LLMs) are increasingly used by researchers and practitioners to generate economic and financial forecasts. In a typical setting, an LLM such as ChatGPT is prompted with textual inputs—such as news headlines or earnings call transcripts—to predict future firm-level outcomes, including stock returns or capital expenditures (e.g., Lopez-Lira and Tang, 2023; Jha et al., 2024). A growing literature finds that LLM-based forecasts often outperform both conventional econometric models and standard machine-learning techniques.

Despite these promising results, their interpretation requires caution. When pre-trained, off-the-shelf LLMs are used, the evaluation effectively occurs in-sample: the model may have been exposed to overlapping or closely related data during training, making it difficult to distinguish genuine reasoning from simple recall. This chance of lookahead bias is especially high when forecasting variables such as stock returns that are widely reported in public sources likely included in the model's training corpus.

Several prompting strategies attempt to mitigate this issue, such as masking firm identifiers, anonymizing contextual information, or limiting access to temporal cues (Glasserman and Lin, 2023; Sarkar, 2024; Engelberg et al., 2025; Wu et al., 2025). Yet their effectiveness remains mixed. Ideally, one would evaluate forecasts strictly on out-of-sample data unseen by the LLM. However, for widely used large-scale models, the available out-of-sample horizon is short, limiting statistical power. Retraining such models on year-by-year corpora, a clean alternative, is computationally prohibitive at current scales.

An additional complication, often overlooked, is that lookahead bias is task-specific rather than an inherent characteristic across LLMs. As shown by Carlini et al. (2023), memorization depends systematically on factors such as the nature and public visibility of the input text and target variable, model size and architecture, and prompt design. These variations underscore the need for a systematic and generalizable test to detect lookahead bias in LLM forecasts on a case-by-case basis.

To address this challenge, we develop a statistical test that determines whether an LLM's forecasts arise from genuine reasoning or from training data leakage. Building on state-of-the-art membership inference attack (MIA) techniques, the proposed method is cost-efficient and requires neither model retraining nor access to proprietary training data. It offers researchers and practitioners a practical diagnostic tool for assessing the reliability

and real-world validity of LLM-generated forecasts.

We first introduce the Lookahead Propensity (LAP), a measure of how likely the text in a prompt appeared in an LLM's training data. For each prompt, we compute the model-assigned probability of every token conditional on its preceding tokens. We define LAP as the mean token probabilities of the bottom K% of tokens, i.e., those tokens with the lowest predicted probabilities. This focus on uncommon tokens is key: frequent words such as "the" or "and" are assigned high probabilities regardless of prior exposure, whereas rare tokens carry more information about whether the text was previously seen. The intuition is that unseen prompts tend to contain more low-probability (outlier) tokens under the model, whereas seen prompts are less likely to have such extreme outliers, and those tokens instead receive higher probability. Accordingly, a higher LAP indicates greater model familiarity and a higher likelihood of training-data overlap.[1]

The LAP corresponds directly to the MIN-K% PROB statistic developed in the membership inference attack (MIA) literature (e.g., Shokri et al., 2017; Carlini et al., 2021, 2022; Shi et al., 2024; Cheng et al., 2024). The MIA literature proposes various measures to detect whether a given text was included in an LLM's training data; among them, MIN-K% PROB statistic stands out as a robust and benchmark-level baseline, demonstrating consistently superior performance across reported evaluation settings (Shi et al., 2024). It has been successfully applied in several LLM engineering contexts, including detecting copyrighted material and auditing the effectiveness of machine-unlearning procedures. See Section 2.1 for a summary. Our contribution lies in applying this measure to prompt inputs used for economic forecasting. We develop a formal statistical test to detect lookahead bias in LLM-based forecasts, which is a central concern for empirical economists.

In Section 2.2, we formally develop our test in an econometric framework. The main proposition implies that if forecast accuracy is correlated with LAP, such a relationship indicates lookahead bias rather than genuine predictive reasoning. Let us first illustrate the intuition using the following example. We prompt the model with:

> *"Here is a piece of news on July 28, 2020: Kodak Triples on Loan to Make Covid-19 Drug Ingredients. Do you think this is good or bad for the stock price in the short term?"*

Without access to future information, the LLM would need to reason about how govern-

[1] Following Shi et al. (2024), we set K% as 20%.

ment pharmaceutical contracts might influence firm value. However, if its training corpus contains subsequent coverage of the same event, the task effectively becomes one of recall rather than reasoning. Indeed, on the following day, July 29, a major news outlet published the headline: "Kodak's stock rose so fast it tripped 20 circuit breakers in a single day." The article body reported:

> *"On Tuesday, President Donald Trump announced the company would receive a $765 million loan to launch Kodak Pharmaceuticals ... Following a more than 200% jump in Tuesday trading, the rally continued on Wednesday and the shares ended up 318%."*

When presented with the July 28 headline, the LLM may recall textual patterns from the July 29 report, such as "shares skyrocketed," "tripped 20 circuit breakers," or "318 percent increase", and consequently infer that the news is positive for the stock price. In this instance, what appears to be predictive ability actually reflects information recall from the training data rather than reasoning over unseen information. Conceptually, a model that achieves high predictive accuracy only when the prompt contains previously seen text is analogous to a student excelling only on exam questions encountered during practice. This pattern indicates memorization rather than genuine comprehension. A high LAP of the input text, therefore, indicates that the LLM has likely been exposed to the realized outcome.

In our econometric framework, the LLM's prediction, denoted by $\hat{\mu}$, consists of two indistinguishable components: genuine reasoning and potential memorization. A standard forecast-accuracy regression of the realized outcome Y on $\hat{\mu}$ cannot disentangle these sources of predictability. Our proposed test augments the regression with an interaction term between $\hat{\mu}$ and LAP. Given that LAP serves as a validated proxy for memorization in the MIA literature, we prove that a positive coefficient on the interaction term is equivalent to the presence of lookahead bias. In other words, if forecast accuracy systematically increases as the input text becomes more familiar to the model, the predictive signal is driven at least partially by memory rather than reasoning.

In Section 3, we implement the LAP test on two forecasting exercises using Llama-3.3, an open-source LLM released by Meta in December 2024. The first exercise adopts the setup in Lopez-Lira and Tang (2023), which uses firm-specific news headlines to predict next-day stock returns. While Lopez-Lira and Tang (2023) focus primarily on the period after the model's knowledge cutoff date, we replicate their analysis over the in-sample period for the

purposes of this study. Our results indicate that memorization substantially amplifies the apparent predictive power of LLM-generated forecasts. Specifically, the baseline regression shows that a one–standard-deviation increase in LLM prediction predicts a 0.197% higher next-day return. By comparison, a one–standard-deviation increase in LAP increases the marginal effect of LLM prediction on next-day stock return by 0.077%, which is about 37% of the standalone LLM effect in the baseline test. Moreover, the stronger predictability observed among small-cap stocks is also largely due to LAP amplification, reflecting stronger lookahead bias.

The amplification effect remains robust when controlling for two confidence measures: the first-token conditional probability of the LLM's response (Chen et al., 2024) and the model's self-reported confidence level. This robustness indicates that memorization operates through a mechanism distinct from the model's internal measure of predictive certainty.

Finally, we conduct the LAP test in a genuinely out-of-sample period using Llama-2, from September 2023 to December 2024. In the period after the release date, no lookahead bias should arise. We first show that the coefficient on the interaction between LAP and the model's prediction indeed becomes statistically insignificant. Next, we conduct a bootstrap analysis. The bootstrap distribution based on out-of-sample data is clearly separated from the in-sample estimate, resulting in a one-sided bootstrap p-value of 0.033 for the in-sample interaction term. These findings further suggest that the in-sample predictability is largely driven by look-ahead bias.

Our second forecasting exercise replicates the analysis in Jha et al. (2024), which uses firms' earnings conference call transcripts to predict subsequent capital expenditures. Consistent with our previous findings, LAP significantly amplifies the relationship between LLM predictions and future investment. The model exhibits stronger predictive performance when analyzing transcripts that contain linguistic patterns familiar from its training data, suggesting that apparent foresight arises at least in part from memorization rather than genuine inference. A one–standard-deviation increase in LLM prediction is associated with a 0.324% higher future two-quarter capital expenditure ratio (scaled by total assets). By comparison, a one–standard-deviation increase in LAP increases the marginal effect of LLM prediction on future CapEx ratio by 0.149%, which is about 19% of the standalone LLM effect in the baseline.

The two exercises demonstrate that at least part of the LLM's apparent predictive power

stems from memorization rather than genuine reasoning, underscoring the need for caution when applying LLMs to economic forecasting. However, this finding does not imply that forecasts by LLMs should be avoided altogether. The extent of lookahead bias is task-specific, not a universal property of LLMs. It depends on factors such as the nature of the input text and target variable, model architecture, and prompt design. The LAP measure and the statistical test we propose provide a simple, cost-effective diagnostic for identifying cases in which model outputs are influenced by memorization rather than analytical reasoning. This distinction is crucial for ensuring the validity of inferences drawn from LLM-based forecasts, especially in backtesting exercises that rely on historical data potentially present in the model's training corpus. As LLMs become increasingly integrated into empirical finance and economic research, systematically distinguishing between memory and reasoning will be essential for enhancing the credibility of their predictions.

Related Literature

A growing body of work applies LLMs to extract economically meaningful signals from financial text. Some studies use corporate disclosures and earnings calls as model inputs to predict firm actions (e.g., Cao et al., 2023; Jha et al., 2024). Others apply LLMs to financial news to forecast stock returns (e.g., Chen et al., 2022; Lopez-Lira and Tang, 2023) and macroeconomic variables (e.g., Hansen and Kazinnik, 2024; Bybee, 2023). Chen et al. (2024) analyze how LLMs' conditional probabilities affect model predictability.

Meanwhile, recent studies have highlighted the potential for lookahead bias in LLM-based predictions (e.g., Glasserman and Lin, 2023; Sarkar and Vafa, 2024). Lopez-Lira et al. (2025) detect LLM memorization by comparing the model's recalled economic variables with their actual values. They find that LLMs can sometimes exactly reproduce historical data from their training period, suggesting that strong forecast performance on pre-cutoff data may reflect memorization rather than reasoning.

Several studies also explore strategies to mitigate such bias. Glasserman and Lin (2023) and Engelberg et al. (2025) introduce an entity-neutering prompting approach, where identifying details such as firm names and dates are removed by LLMs to reduce recognition while retaining informational content. Another line of work develops self-developed models trained under controlled information sets (e.g., Sarkar, 2024; He et al., 2025). For example, He et al. (2025) train leak-free language models—ChronoBERT (149M) and ChronoGPT

(1.5B)—using data available only up to each year-end, and confirm that the findings of Chen et al. (2022) involve only modest look-ahead bias.

Our study complements the existing literature by taking a different approach. Rather than mitigating lookahead bias, we develop a statistical test to assess its likelihood and magnitude. Since lookahead bias in LLM predictions is inherently task-specific, our test provides a more generalizable and adaptable diagnostic tool. Moreover, the proposed LAP test is cost-efficient, requiring neither model retraining nor access to proprietary training data.

This paper also builds on the engineering literature on memorization in language models. Carlini et al. (2021) demonstrate that LLMs can memorize portions of their training data, complicating applications that require strict train–test separation or date-specific cutoffs. Subsequent studies, including Carlini et al. (2023) and Cheng et al. (2024), show that token-level likelihoods and perplexity are effective indicators of memorization. The MIA literature—reviewed comprehensively in Section 2.1—has further developed a suite of quantitative measures and validation procedures to determine whether a given text is likely part of an LLM's training corpus. Our contribution lies in extending these insights by applying MIA techniques to develop a statistical test for detecting lookahead bias in LLM-based economic forecasts.

2 Lookahead Bias Detection

2.1 Lookahead Propensity (LAP)

We introduce the statistical properties of language models, the MIA literature, and LAP construction in this subsection.

Language Model A large language model (LLM) is a probabilistic framework that predicts the next token in a text sequence.[2] Given a sequence of observed tokens $w_{\leq n-1} := (w_1, w_2, \ldots, w_{n-1})$, an LLM parameterized by θ estimates the likelihood of the next token w_n through the conditional probability

$$P_\theta\left(w_n \mid w_{\leq n-1}\right) = P_\theta(w_n \mid w_1, w_2, \ldots, w_{n-1}).$$

[2]In natural language processing (NLP), a *token* is the smallest unit of text that carries meaning, such as a word, subword, or character, depending on the tokenization scheme.

LLMs learn these conditional probabilities by training on massive text corpora, adjusting internal parameters to maximize the likelihood of observed sequences.

Through repeated exposure to billions of examples, the model internalizes statistical patterns—strengthening associations between words, phrases, and events that frequently co-occur. This process does not involve "understanding" in the human sense; rather, LLMs build a compressed representation of co-occurrence patterns across language. When presented with a prompt, the model retrieves relevant patterns from this learned distribution to generate contextually coherent continuations.

Critically for our analysis, this learning process absorbs not only contemporaneous associations but also retrospective narratives written after the occurrence of the event of interest, creating the potential for lookahead bias in applications that treat model outputs as real-time information.

LLM Forecast and Lookahead Bias Training data often contain both original news and subsequent articles describing market reactions, allowing models to learn events and outcomes together. In the Kodak example from Section 1, training data likely include both the loan announcement and next-day coverage of the stock surge. The model thus learns not only that "a loan announcement happened" but also that "it was followed by a sharp stock increase." When the model encounters similar news later, it may effectively already "know" what happens next. This creates a kind of lookahead bias, where predictions mirror memorized outcomes rather than genuine reasoning.

When prompted with a headline consisting of N tokens $(w_1, \ldots, w_N)$, the model computes

$$P_\theta(\text{prediction} \mid \text{headline}) = P_\theta(w_{N+1} \mid w_{\leq N})$$

where w_{N+1} corresponds to the highest probability category ("positive," "negative," or "neutral").[3] These probabilities reflect statistical associations from training rather than causal reasoning. For the Kodak headline "Kodak Triples on Loan to Make Covid-19 Drug Ingredients" (Figure I), the model's prediction likely reflects memorized event-outcome pairs from its July 2020 training data.

Because media coverage of an event and its subsequent outcome typically occur in close succession, the inclusion of event-related articles in an LLM's training corpus likely coin-

[3] Since we select the maximum probability token at each step, output is deterministic given the same input.

cides with the inclusion of outcome-related coverage. This time-period overlap introduces a form of lookahead bias, whereby the model's apparent ability to anticipate future outcomes may partly arise from exposure to post-event information during training. Building on this intuition, we can infer lookahead bias from the model's memorization of event-specific news headlines, quantified using our LAP measure introduced below.

MIA Techniques Membership inference attacks (MIAs), introduced by Shokri et al. (2017), formalize the task of determining whether a sample was included in a machine learning (ML) model's training set through a shadow-model framework. Subsequent studies such as Yeom et al. (2018) refined attack methods and evaluation metrics. Later, Carlini et al. (2022) framed MIA as a hypothesis-testing problem using likelihood-ratio principles, offering a more rigorous theoretical foundation.

The rise of LLMs has spurred growing interest in testing whether specific texts were included in a model's pre-training corpus, a setting distinct from classical MIAs developed for conventional ML models. This interest is driven both by the opaque nature of LLM training data and by concerns over the potential extraction of copyrighted or sensitive content (e.g., Carlini et al., 2021).

Shi et al. (2024) advance this line of work by formally defining reference-free pretraining-data detection problem and introducing the time-split WIKIMIA benchmark for reliable validation. A key insight motivating their approach is that not all tokens' conditional probabilities provide equal information about memorization. Common words such as "the," "is," or "to" appear frequently across all texts and thus have high probabilities regardless of whether the model has seen the specific content before. These high-frequency tokens add noise rather than signal to detection efforts.

Building on this insight, Shi et al. (2024) propose MIN-K% PROB, a simple, training-free detection score for assessing whether a given text has been memorized by an LLM. The method focuses on the most informative tokens—those assigned unusually low probabilities by the model. Formally, for a passage with n tokens, let the LLM assign each token a conditional probability p_i. MIN-K% PROB sorts tokens by p_i and computes the average log-probability of the bottom K%—the least likely tokens under the model. The hyperparameter K determines the sensitivity of the measure to low-probability "outlier" tokens.

Empirically, $K = 20$ yields the most robust performance. Evaluated on WIKIMIA across

multiple large language models, Shi et al. (2024) show that MIN-K% PROB attains an average AUC of 0.72, significantly higher than alternative MIA measures.[4] On an additional copyrighted-book validation test, it achieves an AUC of 0.88. Also, its detection performance generally improves for larger models and longer passages. Therefore, we adopt MIN-K% PROB with $K = 20$ as our primary detection method given its strong empirical performance.

Construction of LAP We now formally define LAP (i.e., MIN-K% PROB). Consider a language model with parameters θ, when presented with a prompt $w = (w_1, \ldots, w_N)$, LAP is computed as,

$$\text{Lookahead Propensity}(w, K) \coloneqq \exp\left(\frac{1}{|S_K|} \sum_{t \in S_K} \log P_\theta(w_n \mid w_{\leq n-1}) \right), \tag{1}$$

where $P_\theta(w_n \mid w_{\leq n-1})$ is the conditional probability assigned by model θ to token w_n given its preceding context $w_{<n} \coloneqq (w_1, \ldots, w_{n-1})$, and $S_K \subset \{1, \ldots, N\}$ denotes the subset containing the lowest K% of tokens ranked by $P_\theta(w_n \mid w_{\leq n-1})$, and $K = 20$. The input w encompasses any text provided to the model, including prompts, headlines, or full news articles, as exampled in Figures I and II.

2.2 Econometric Framework

We formalize lookahead bias through a contamination model where predictions incorporate future information via memorization mechanisms. Formally, following Sarkar and Vafa (2024), lookahead bias manifests when a model's prediction $\hat{\mu}_t = \hat{\mu}(X_t)$ for time $t+1$ violates the orthogonality condition:

$$\text{Cov}(\hat{\mu}_t, \varepsilon_{t+1}) \neq 0. \tag{2}$$

Consider a standard forecasting environment with the following data-generating process:

$$Y_{t+1} = \mu(X_t) + \varepsilon_{t+1}, \tag{3}$$

[4] AUC (area under the ROC curve) measures how well a detector distinguishes members from non-members. It equals the probability that a random member gets a higher score than a random non-member. An AUC of 0.5 means no better than guessing; 1.0 means perfect detection.

where Y_t denotes the realized outcome and X_t observable information at time t (e.g., news headlines, earnings call transcripts), $\mu(X_t) = \mathbb{E}[Y_{t+1} \mid X_t]$ is the true conditional expectation, and $\varepsilon_{t+1} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ represents future innovations unpredictable given the information set $\mathcal{M}_t$ available at time t, with $\mu(X_t) \perp \varepsilon_{t+1}$.

Definition 1 (Lookahead Bias Contamination)**.** *When a LLM suffers from lookahead bias, its predictions take the form:*

$$\hat{\mu}_t = \mu(X_t) + L_t \varepsilon_{t+1} \tag{4}$$

where L_t measures memorization strength.

The key implications of our framework are as follows. First, the contamination term $L_t \varepsilon_{t+1}$ violates the orthogonality condition in Equation (2), thereby introducing lookahead bias whenever $L_t > 0$. Specifically, $\text{Cov}(L_t \varepsilon_{t+1}, \varepsilon_{t+1}) = L_t \text{Var}(\varepsilon_{t+1}) > 0$ implies that any positive L_t induces mechanical correlation between the model prediction and the irreducible error.

Second, the magnitude of this bias scales with the degree of LLM memorization L_t. LLMs trained with greater exposure to future data implicitly encode information about later-period content, leading to stronger deviations from the true conditional process even when no explicit future reference is present.

Third, the resulting predictor exhibits distorted accuracy, since its expected squared prediction error

$$\mathbb{E}[(\hat{\mu}_t - Y_{t+1})^2] = \text{Var}(\varepsilon_{t+1})[1 - L_t]^2,$$

falls mechanically as L_t increases.

Finally, two boundary cases clarify the mechanism: when $L_t = 0$ for all L_t, pretraining restricted to past data eliminates bias; whereas as $L_t \to 1$, complete memorization of future information produces full contamination.

Based on our earlier discussion, we proxy L with LAP defined in Equation (1). To detect lookahead bias, we propose testing the significance of the interaction term in the regression:

$$Y_{t+1} = \beta_1 \hat{\mu}_t + \beta_2 L_t + \beta_3 (L_t \times \hat{\mu}_t) + \epsilon_{t+1} \tag{5}$$

Proposition 1 (Detection Statistic)**.** *By the Frisch–Waugh–Lovell (FWL) theorem,*

$$\beta_3 = \frac{Cov(\tilde{y}, \widetilde{L\hat{\mu}})}{\mathrm{Var}(\widetilde{L\hat{\mu}})},$$

where $\tilde{y}$ *and* $\widetilde{L\hat{\mu}}$ *are the residuals from projecting* Y_{t+1} *and* $L\hat{\mu}$ *onto* $(\hat{\mu}, L)$*, respectively. Hence,*

$$\beta_3 > 0 \iff Cov(\tilde{y}, \widetilde{L\hat{\mu}}) > 0.$$

The partial covariance admits the structural representation

$$Cov(\tilde{y}, \widetilde{L\hat{\mu}}) = \mathbb{E}\big[L^2 \,\mathrm{Var}(\varepsilon \mid \hat{\mu}, L)\big]. \tag{6}$$

Therefore,

$$\beta_3 > 0 \quad \textit{iff} \quad \mathbb{E}\big[L^2 \,\mathrm{Var}(\varepsilon \mid \hat{\mu}, L)\big] > 0,$$

which holds whenever L *is nonzero on a set of positive probability and* $\mathrm{Var}(\varepsilon \mid \hat{\mu}, L) > 0$ *on that set. In particular, if* $L \geq 0$ *a.s. and* $\Pr(L > 0) > 0$*, then* $\beta_3 > 0$*. If* $L = 0$ *a.s., then* $\beta_3 = 0$*.*

Proof: See Appendix B.1.

Interpretation Equation (6) implies that the covariance term is strictly positive whenever $L > 0$, indicating that lookahead bias arises whenever $\beta_3 > 0$. The magnitude of this covariance scales jointly with the intensity of memorization L, and it vanishes when $L = 0$, corresponding to the absence of lookahead bias. This formulation provides a theoretically grounded test statistic for empirically detecting such bias. The result also clarifies why common mitigation strategies such as prompt engineering or identifier masking are ineffective: they do not eliminate the structural contamination term $L_t \varepsilon_{t+1}$ that embeds future information in LLM outputs.

3 Result

3.1 Data

Stock Market Information Daily stock returns, open prices, and close prices are obtained from the Center for Research in Security Prices (CRSP), covering all common stocks listed

on the New York Stock Exchange (NYSE), the National Association of Securities Dealers Automated Quotations (NASDAQ), and the American Stock Exchange (AMEX). We restrict the sample to common stocks with a share code of 10 or 11, consistent with prior studies. We obtain financial statement variables from Compustat.

Stock News Headlines Data on news headlines are collected via web scraping from Bloomberg News. We select Bloomberg News because it is part of Bloomberg L.P., a leading global financial information provider whose real-time news service is updated more frequently than traditional outlets. We collect the news headlines for all CRSP-listed companies in the sample period and match them based on company names and ticker symbols. The final dataset consists of 91,361 news headlines covering 1,587 unique companies from January 2012 to December 2023.

Earnings Call Transcripts We obtain data on firms' earnings call transcripts from Thomson Reuter's StreetEvents database. Then, we merge them with CRSP and Compustat data using firm tickers and corresponding call dates, resulting in 135,541 matched observations. After excluding records with missing values in the key variables, our final sample consists of 74,338 firm-quarter observations from 3,897 unique U.S. publicly listed firms between 2006 and 2020.

3.2 LLM Setup and Prompt Design

Our analysis adopts Llama-3.3, an open-source model released by Meta in December 2024. We choose Llama-3.3 over proprietary alternatives such as OpenAI's API for two key methodological reasons.

Most importantly, Llama-3.3 provides access to prompt token probabilities, which are essential for computing our LAP measure. While OpenAI exposes token probabilities for generated output tokens, it does not provide access to prompt token probabilities, making it unsuitable for our research design.

Additionally, the open-source nature of Llama-3.3 ensures long-term replicability. Model checkpoints, which are saved snapshots of trained model parameters, can be freely downloaded from platforms such as HuggingFace. This allows any researcher to reproduce our analysis using the identical December 2024 version. In contrast, proprietary API providers

may update or deprecate models over time (as OpenAI has done with ChatGPT 3.5), preventing future replication using the exact model version.

For the stock news exercise, we follow the prompt design in Lopez-Lira and Tang (2023). To ensure robustness in our analysis, we also instruct the LLM to provide a confidence score alongside its prediction. The prompt explicitly instructs the model to predict whether the news is "good", "bad", or "neutral" for the stock price of the mentioned company,[5] accompanied by a confidence score and a brief explanation, as shown in Figure I. This structured format is intended to enhance interpretability, allowing for clear numerical indicators for subsequent analysis.

For each headline, the model's response is parsed and mapped to a numerical score, where good is mapped to $+1$, neutral is mapped to 0, and bad is mapped to -1. The confidence score is directly recorded from the LLM's output, while the explanation is stored for interpretive analysis.

For the earnings call exercise, we adopt the approach of Jha et al. (2024), similarly incorporating confidence scores and limiting the input to the first 500 words of each transcript, as shown in Figure II. The prediction signal generated by the LLM for firm i in quarter q is based on its earnings call transcript. From the LLM prediction on earnings call data, the variable takes values of -1, -0.5, 0, 0.5, or 1, indicating whether future capital expenditures are expected to significantly decrease, slightly decrease, not change, slightly increase, or significantly increase before they are realized.

We use vLLM to run the Llama-3.3 checkpoint locally, enabling efficient batched inference on GPU. For each headline prompt, we compute the LAP measure by requesting token-level log probabilities via `prompt_logprobs = 1`. We configure inference with `do_sample = False` to ensure deterministic token selection and full reproducibility across runs.

3.3 Prompt News Headlines to Predict Stock Returns

In this exercise, we generally follow the approach of Lopez-Lira and Tang (2023), with a few differences. First, their study evaluates model performance in a post–knowledge-cutoff

[5] We also conduct robustness tests by replacing "stock price" with "stock return", as returns more directly reflect changes in market expectations. This aligns with Baker and Wurgler (2006) who showed that sentiment-driven mispricing in certain stock categories creates predictable return patterns—underpricing during low sentiment periods leads to higher subsequent returns, while overpricing during high sentiment leads to lower returns. The results remain qualitatively similar.

period of ChatGPT-4, which can be interpreted as an out-of-sample setting, whereas our tests are conducted in a pre–knowledge-cutoff period for this purpose of our study. Second, we use news headlines from Bloomberg, while they collect headlines from multiple media sources. Third, their analysis relies on ChatGPT-4 to classify each headline as good (+1), bad (−1), or neutral (0); we adopt the same classification scheme but instead use Llama-3.3 for the reasons discussed above. As shown later, the choices on news source and model do not materially affect the results, as we successfully replicate their main findings within our out-of-sample analysis.

For simplicity, we use stock market close time 4 p.m. as the cutoff to classify headlines on day t. Specifically, if a headline is published before 4 p.m., it is considered news on day t and will link it to the next-day stock return; otherwise, it is treated as news on day $t+1$.

We begin with the in-sample analysis, as the LAP measure is designed to detect potential memorization within the model's own training period, spanning January 2012 to December 2023. We then repeat the analysis on the out-of-sample data as a placebo test. In contrast, due to a different research focus, Lopez-Lira and Tang (2023) examine an out-of-sample period from October 2021 to May 2024, consistent with ChatGPT-4's reported knowledge cutoff of September 2021.

Following the discussion in Section 2.2, the baseline regression is specified as follows:

$$r_{i,t+1} = \alpha_i + \beta_t + \gamma \cdot \text{LLM}_{i,t} + \delta \cdot (\text{LLM}_{i,t} \times \text{LAP}_{i,t}) + \eta \cdot \text{LAP}_{i,t} + u_{i,t+1}, \tag{7}$$

where $r_{i,t+1}$ is the stock return of firm i on day $t+1$ (in %), $\text{LLM}_{i,t} \in \{-1, 0, +1\}$ derived from the LLM's evaluation of news headlines, and $\text{LAP}_{i,t}$ is the lookahead propensity for the news headline as specified in equation 1. Following Lopez-Lira and Tang (2023), we includes firm and date fixed effects. Panel A of Table I shows summary statistics of main variables.

Table II presents our results. The specification in Column (1) does not include the interaction term and thus replicates the main finding of Lopez-Lira and Tang (2023). It shows that LLMs can predict returns: a one–standard-deviation increase in the LLM prediction measure predicts a 0.197% higher next-day return. Column (2) introduces our main test. The coefficient before the interaction term $\text{LLM}_{i,t} \times \text{LAP}_{i,t}$ is positive and highly significant. This means that the LLM's predictive power increases dramatically for high-LAP headlines. Economically, a one–standard-deviation increase in LAP increases the marginal effect of

LLM on next-day stock return by 0.077%, which is about 37% of the standalone LLM effect shown in Column (1). The coefficient of LLM itself become insignificant.

Lopez-Lira and Tang (2023) also find that LLM-based predictions exhibit stronger forecasting power for small firms. They attribute this pattern to the relatively less efficient price discovery in small-cap stocks, which tend to adjust more slowly to new information. We revisit this result in Table III, Column (1), by adding an interaction term between LLM and Small, a dummy variable equal to one for firms in the bottom quintile of market capitalization, into the our main regression. The coefficient on the interaction term is positive, consistent with their findings. In Column (2), we further introduce a triple interaction term, LLM × Small × LAP, to examine whether the stronger predictability for small-cap stocks is also driven by lookahead bias, as specified below:

$$\begin{aligned} r_{i,t+1} \quad = \quad & \alpha_i + \beta_t + \gamma_1 \text{LLM}_{i,t} + \gamma_2 \text{Small}_{i,t} + \gamma_3 (\text{LLM}_{i,t} \times \text{Small}_{i,t}) \\ & + \delta_1 \text{LAP}_{i,t} + \delta_2 (\text{LLM}_{i,t} \times \text{LAP}_{i,t}) + \delta_3 (\text{LAP}_{i,t} \times \text{Small}_{i,t}) \\ & + \delta_4 (\text{LLM}_{i,t} \times \text{LAP}_{i,t} \times \text{Small}_{i,t}) + u_{i,t+1}. \end{aligned} \tag{8}$$

The results confirm this suspect: the coefficient on the triple interaction term is positive and statistically significant. Moreover, once the triple interaction is included, the coefficient on LLM × LAP becomes negative.

These patterns are consistent with lookahead bias. The apparent predictability of the LLM increases with higher LAP values, indicating that its forecasts are more accurate when the input text is familiar to the model.

Compare to LLM Prediction Confidence Recent studies have examined how an LLM's self-reported or inferred confidence interacts with its forecast accuracy. For example, using a similar financial-news and stock-return setting, Chen et al. (2024) show that the conditional probability of an LLM's response contains economically meaningful information and improves predictive performance. Although these confidence measures are conceptually distinct from our LAP metric, it is worthwhile to test empirically whether they capture related mechanisms.

Specifically, we incorporate two confidence measures into our analysis: (1) the first-token conditional probability in the LLM response $p(w_{i,N+1} \mid w_{i,\leq N})$, which reflects the

model's inferred belief in its prediction given the wording of the prompt (Chen et al., 2024), and (2) the model's self-reported confidence level, $\text{Confidence}_{i,t}$. We then conduct horse-race regressions between these confidence measures and LAP to assess whether model confidence and memorization propensity explain overlapping or distinct components of LLM forecasting performance.

Table IV shows the results. In Column (1), we include LLM prediction, $p(w_{i,N+1} \mid w_{i,\leq N})$, and their interaction term. We find that the coefficient before the interaction term is significantly positive, consistent with findings of Chen et al. (2024). In Column (2), we add LAP and LLM × LAP. The main finding remains robust: the coefficient on LLM × LAP is still significantly posistive with a similar magnitude to that reported in Table II. The effect of $p(w_{i,N+1} \mid w_{i,\leq N})$ on LLM do not change materially either. In Columns (3) and (4), we repeat the analysis by using model's self-reported confidence, and the results are similar.

Taken together, the result suggests that LAP and model confidence capture distinct components of LLM forecasts.

Out-of-Sample and Bootstrap Analysis Finally, we conduct a placebo test to validate our proposition. Specifically, we repeat the analysis using news headlines released after the model's release date, during which forecasts should not be subject to lookahead bias. In this out-of-sample period, the LAP measure should capture only random noise rather than any memorization of the input text. Accordingly, in the placebo regression, the interaction term between LAP and LLM is expected to yield an insignificant coefficient, consistent with the absence of lookahead bias once the model no longer has access to training-period information.

To implement this idea, we use Llama-2 instead of Llama-3.3, which was released in December 2024 and therefore offers too short a horizon for our test. Llama-2 was released in July 2023 and updated with some of the files in Aug 2023, providing a longer evaluation window. To be conservative, we use the model's release date—rather than its claimed knowledge cutoff—as the starting point for our out-of-sample period. Thus, the out-of-sample period is from September 2023 to December 2024. Appendix Table A.1 summarizes the corresponding in-sample and out-of-sample periods for each model.

To ensure comparability across in-sample and out-of-sample periods, we standardize all variables separately within each period and re-estimate the same set of regressions. Table

V reports the results. Columns (1) and (2) replicate our main in-sample findings, which are qualitatively similar to those obtained using Llama-3.3. LLM predictions are positively correlated with next-day stock returns, and this relationship is stronger when LAP is high.

In Columns (3) and (4), we re-estimate the regressions using the out-of-sample data. In Column (3), the coefficient on LLM is significant, consistent with the findings in Lopez-Lira and Tang (2023). As shown in Column (4), the coefficient on the interaction term $\text{LLM}^{\text{std}} \times \text{LAP}^{\text{std}}$ is negative and statistically insignificant, consistent with the absence of lookahead bias during the out-of-sample period.

Next, we conduct a bootstrap analysis. We use the pairs bootstrap approach (e.g., Efron and Tibshirani, 1994). Specifically, we repeatedly sample observations from the out-of-sample data with replacement for 10,000 repetitions and re-estimate the regression in each bootstrap sample. The bootstrap sample size is set equal to that of the out-of-sample period. For each iteration, we record the estimated coefficient on the interaction term between LLM^{std} and LAP^{std}.

Figure III shows the empirical distribution of the bootstrap coefficients. The blue, dotted vertical line indicates the 95th percentile of the empirical bootstrap distribution, while the red, solid vertical line represents the in-sample estimate of the interaction coefficient shown in Column (2). Consistent with our baseline results, the distribution in the out-of-sample placebo period is centered at a negative value. Moreover, the in-sample interaction coefficient lies well outside the 95th percentile of the empirical distribution of the out-of-sample interaction coefficients (with one-sided p-value of 0.033), highlighting a clear difference between the in-sample and out-of-sample distributions.

3.4 Prompt Earnings Call Transcripts to Predict Capex

In our second forecast exercise, we follow the approach of Jha et al. (2024), who use earnings call transcripts to predict firms' capital expenditures two quarters ahead. We generally follow their procedure of data cleaning and variable construction for consistency. In their analysis, each transcript is segmented into 2,000-word chunks and classified by ChatGPT-3.5 into five categories: significantly increase (+1.0), slightly increase (+0.5), no change (0.0), slightly decrease (−0.5), and significantly decrease (−1.0). We use the same classification scheme but replace ChatGPT-3.5 with Llama-3.3. Due to the context limitation of the self-hosted Llama-3.3 model, we restrict each input to the first 512 words of

the transcript, yet we are still able to replicate their baseline results. Our sample period is 2006–2020, identical to theirs. Summary statistics of main variables are reported in Panel B of Table I.[6]

Jha et al. (2024) find that, when prompted with earnings call transcripts, LLMs can effectively predict firms' subsequent CapEx ratios. We examine whether this apparent predictability is driven by memorization, as proxied by LAP. Column (1) of Table VI reproduces their main finding: the LLM-generated prediction score $\text{LLM}_{i,q}$ strongly forecasts future capital expenditure.

Column (2) introduces the interaction term between LLM and LAP. The estimation indicates that the observed predictability is amplified by memorization: the coefficient on the interaction term is positive and highly significant. In terms of economic magnitude, a one-standard-deviation increase in the LLM prediction corresponds to a 0.324% higher CapEx ratio, while a one-standard-deviation increase in LAP raises the marginal effect of LLM by 0.149%—approximately 19% of the standalone LLM effect in Column (1).[7]

The pattern suggests that the model's apparent ability to extract forward-looking information from corporate communications may partly reflect its memory of firms' investment activities. This is not surprising, as articles discussing a company's investments and earnings-call communications are likely included in the model's training corpus.

4 Conclusion

Large language models (LLMs) are increasingly used to generate economic and financial forecasts, often appearing to outperform conventional econometric and machine-learning methods. However, recent studies suggest that this apparent predictive power may partly reflect lookahead bias, as evaluation is effectively conducted in-sample, and the model may have been exposed to overlapping information during training.

We develop a statistical test to distinguish genuine reasoning from training-data leakage in LLM-based forecasts. Building on MIA techniques, we estimate the likelihood that a given prompt appeared in an LLM's pre-training corpus—a measure we term Lookahead

[6]The distribution of LAP can vary substantially with the length of the input text, which explains why the LAP distributions differ markedly across the two exercises.

[7]We were unable to replicate the baseline results using Llama-2, potentially due to limitations in its ability to handle long-context inputs. As a result, we do not include out-of-sample test results for the earnings call exercise.

Propensity (LAP). We show that a positive correlation between LAP and forecast accuracy indicates the presence and magnitude of lookahead bias.

Applying the LAP test to two forecasting tasks—news headlines predicting stock returns and earnings-call transcripts predicting capital expenditures—we find that the predictive power of LLMs partly reflects memorization rather than genuine inference.

Lookahead bias is task-specific, varying with factors such as input visibility, target variable, model scale, and prompt design. Our LAP test provides a cost-efficient, diagnostic tool that requires no model retraining or proprietary data access, offering a practical framework for assessing the validity and reliability of LLM-generated forecasts.

References

M. Baker and J. Wurgler. Investor sentiment and the cross-section of stock returns. *The journal of Finance*, 61(4):1645–1680, 2006.

J. L. Bybee. The ghost in the machine: Generating beliefs with large language models. *arXiv preprint arXiv:2305.02823*, 2023.

S. Cao, W. Jiang, B. Yang, and A. L. Zhang. How to talk when a machine is listening: Corporate disclosure in the age of ai. *The Review of Financial Studies*, 36(9):3603–3642, 2023.

N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021.

N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pages 1897–1914. IEEE, 2022.

N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023.

H. Chen, A. Didisheim, and L. Somoza. Out of the black box: Uncertainty quantification for llms via conditional probabilities. *Available at SSRN*, 2024.

Y. Chen, B. T. Kelly, and D. Xiu. Expected returns and large language models. *Available at SSRN 4416687*, 2022.

J. Cheng, M. Marone, O. Weller, D. Lawrie, D. Khashabi, and B. Van Durme. Dated data: Tracing knowledge cutoffs in large language models. *First Conference on Language Modeling*, 2024.

B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

J. Engelberg, A. Manela, W. Mullins, and L. Vulicevic. Entity neutering. *Available at SSRN*, 2025.

P. Glasserman and C. Lin. Assessing look-ahead bias in stock return predictions generated by gpt sentiment analysis. *arXiv preprint arXiv:2309.17322*, 2023.

A. L. Hansen and S. Kazinnik. Can chatgpt decipher fedspeak? *Available at SSRN 4399406*, 2024.

S. He, L. Lv, A. Manela, and J. Wu. Chronologically consistent large language models. *arXiv preprint arXiv:2502.21206*, 2025.

M. Jha, J. Qian, M. Weber, and B. Yang. Chatgpt and corporate policies. *arXiv preprint arXiv:2409.17933*, 2024.

A. Lopez-Lira and Y. Tang. Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*, 2023.

A. Lopez-Lira, Y. Tang, and M. Zhu. The memorization problem: Can we trust llms' economic forecasts? *arXiv preprint arXiv:2504.14765*, 2025.

S. K. Sarkar. Storieslm: A family of language models with time-indexed training data. *Available at SSRN 4881024*, 2024.

S. K. Sarkar and K. Vafa. Lookahead bias in pretrained language models. *Available at SSRN 4754678*, 2024.

W. Shi, A. Ajith, M. Xia, Y. Huang, D. Liu, T. Blevins, D. Chen, and L. Zettlemoyer. Detecting pretraining data from large language models. *The Twelfth International Conference on Learning Representations*, 2024.

R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

K. Wu, B. Yang, Z. Ying, and D. Zhou. Anonymization and information loss. *arXiv preprint arXiv:2511.15364*, 2025.

S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.

Figure I. Example Prompt and Response for Stock News Analysis

Example Prompt:

Here is a piece of news: *"(2020-07-28) Kodak Triples on Loan to Make Covid-19 Drug Ingredients."*

Do you think this news is good, bad, or neutral for the stock price of *Eastman Kodak Company (KODK)* **in the short term?**

Write your answer as:
{**good/ bad / neutral**}
{**confidence (0-1):**}
{**explanation (less than 25 words)**}

Example Response:

{**good**}
{**1.0**}
{**Loan to produce Covid-19 drug ingredients boosts prospects.**}

Figure II. Example Prompt and Response for Earnings Call Analysis

Example Prompt:

Here is an excerpt from the earnings call transcript of *Oceaneering International Inc. (OII)***:**

Q1 2020 Oceaneering International Inc. Earnings Call
"...We began 2020 with the expectation of marginal growth and improving business fundamentals across all of our segments. And then the COVID-19 pandemic erupted and fueled the further deterioration of the crude oil market fundamentals..."

Do you think the company Oceaneering International Inc. (OII) plans to increase or decrease its capital expenditures over the next year?

Write your answer as:
{**significantly_increase / slightly_increase / no_change / slightly_decrease / significantly_decrease**}
{**confidence (0–1):**}
{**explanation (less than 25 words):**}

Example Response:

{**significantly_decrease**}
{**0.8**}
{**Due to COVID-19 pandemic and crude oil market deterioration.**}

Figure III. Results from Pairs Bootstrap Inference

This figure illustrates the bootstrap distribution of the interaction coefficient β between standardized LLM and LAP, from the specification

$$r_{i,t+1} = \gamma_1 \, \mathrm{LLM}^{\mathrm{std}}_{i,t} + \gamma_2 \, \mathrm{LAP}^{\mathrm{std}}_{i,t} + \beta \left(\mathrm{LLM}^{\mathrm{std}}_{i,t} \times \mathrm{LAP}^{\mathrm{std}}_{i,t} \right) + \mu_i + \lambda_t + u_{i,t+1}.$$

In each bootstrap replication $b = 1, \ldots, 10{,}000$, out-of-sample observations are resampled with replacement, the regression above is re-estimated, and an interaction coefficient $\beta^{(b)}$ is obtained. The histogram plots the empirical distribution of these $\{\beta^{(b)}\}_{b=1}^{10{,}000}$. The red solid vertical line denotes the interaction coefficient estimated from in-sample, while the blue dotted vertical line marks the 95th percentile of the bootstrap distribution.

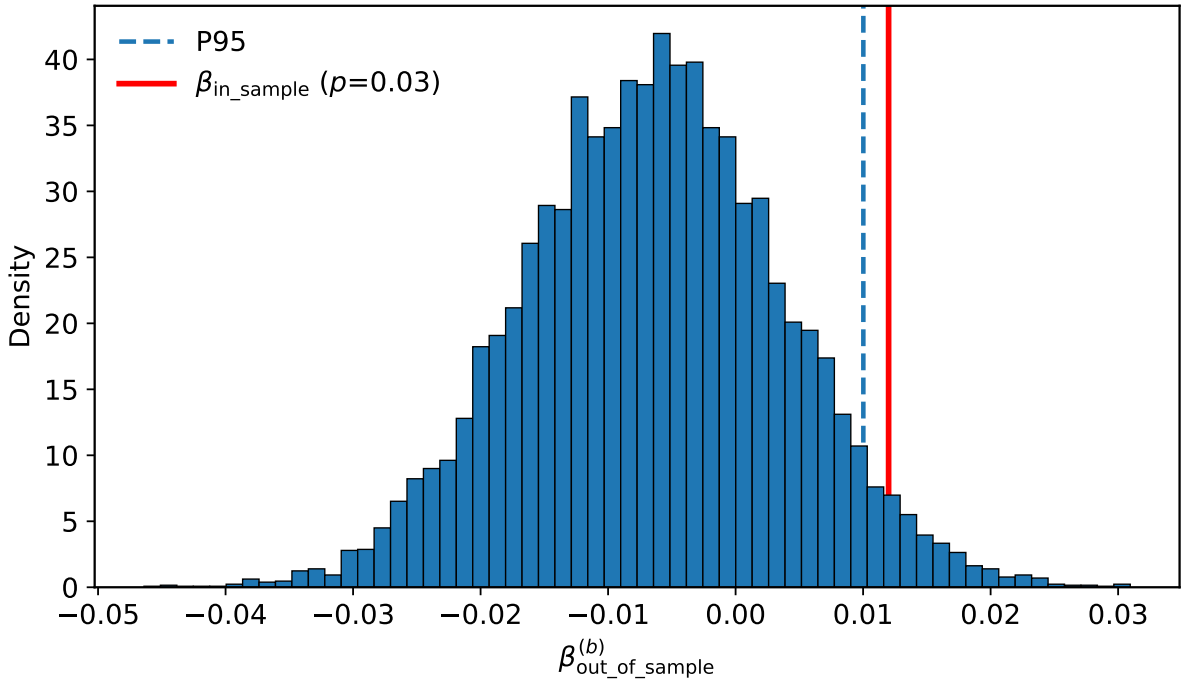

Table I. Descriptive Statistics

Panel A: Stock News Prediction

This table presents descriptive statistics for key variables used in the stock news analysis. Statistics reported include mean, standard deviation (SD), and percentiles (P10, P25, Median, P75, P90).

Variable	Mean	SD	P10	P25	Median	P75	P90	N
LAP ($\times 10^4$)	0.073	0.027	0.041	0.053	0.069	0.089	0.110	91,361
r_{t+1} (%)	0.061	4.301	-2.599	-1.036	0.035	1.136	2.626	91,361
LLM	0.084	0.937	-1.000	-1.000	0.000	1.000	1.000	91,361
$p(w_{N+1} \mid w_{\leq N})$	0.987	0.054	0.988	0.999	1.000	1.000	1.000	91,361
Confidence	0.801	0.073	0.700	0.800	0.800	0.800	0.900	91,361

Panel B: Earnings Call Prediction

This panel presents descriptive statistics for key variables used in the corporate investment prediction analysis. Statistics reported include mean, standard deviation (SD), and percentiles (P10, P25, Median, P75, P90).

Variable	Mean	SD	P10	P25	Median	P75	P90	N
LAP (%)	2.049	1.007	0.947	1.330	1.872	2.573	3.373	74,338
CapEx (%)	2.823	3.967	0.108	0.479	1.427	3.471	7.104	74,338
LLM	0.076	0.407	-0.500	0.000	0.000	0.500	0.500	74,338

Table II. Regression of Next Day Returns on LLM Prediction and LAP

This table reports regression results from the specification

$$r_{i,t+1} = \alpha_i + \beta_t + \gamma \cdot \mathrm{LLM}_{i,t} + \delta \cdot (\mathrm{LLM}_{i,t} \times \mathrm{LAP}_{i,t}) + \eta \cdot \mathrm{LAP}_{i,t} + u_{i,t+1},$$

where $r_{i,t+1}$ is the next-day stock return for firm i at time t, measured in percentage points. The variable $\mathrm{LLM}_{i,t}$ represents the LLM-generated investment signal, taking values of -1, 0, or $+1$, indicating negative, neutral, or positive predictions, respectively. The measure $\mathrm{LAP}_{i,t}$ is computed by averaging the model's token probabilities for the hardest-to-predict $K = 20\%$ of prompt tokens. Column (1) reports results from a specification that excludes the interaction term between $\mathrm{LLM}_{i,t}$ and $\mathrm{LAP}_{i,t}$. Column (2) adds this interaction term to evaluate whether the predictive power of the LLM signal varies with LAP. All regressions include firm and time fixed effects, and standard errors are clustered by date. The sample includes all U.S. common stocks with at least one news headline covering the firm from January 2012 to December 2023.

	(1)	(2)
	$r_{i,t+1}$	$r_{i,t+1}$
$\mathrm{LLM}_{i,t}$	0.210***	0.00117
	(12.24)	(0.03)
$\mathrm{LAP}_{i,t}$		-1.297***
		(-2.61)
$\mathrm{LLM}_{i,t} \times \mathrm{LAP}_{i,t}$		2.866***
		(4.86)
Firm FE	Yes	Yes
Date FE	Yes	Yes
R^2	0.179	0.180
N	91,361	91,361

Table III. Regression of Next-Period Returns on LLM Prediction, LAP, and Firm Size

This table reports regression results examining how the predictive power of LLM-based prediction varies with LAP and firm size. The variable $\text{LLM}_{i,t}$ denotes the prediction score generated by the model for firm i at time t, taking values of -1, 0, or $+1$ to indicate negative, neutral, or positive prediction, respectively. The measure $\text{LAP}_{i,t}$ is computed by averaging the model's token probabilities for the hardest-to-predict $K = 20\%$ of prompt tokens. The indicator $\text{Small}_{i,t}$ equals one if the firm is in the bottom quintile of market capitalization based on the previous day's market cap, and zero otherwise. Column (1) reports a specification that includes the LLM score, the small-firm indicator, and their interaction to assess whether the return predictability of the LLM signal differs between small and large firms. Column (2) adds the interaction between $\text{LLM}_{i,t}$ and $\text{LAP}_{i,t}$, as well as the corresponding triple interaction with $\text{Small}_{i,t}$, to evaluate whether the predictive power of the LLM signal varies jointly with LAP and firm size. All regressions include firm and time fixed effects, and standard errors are clustered by date. The sample includes all U.S. common stocks with at least one news headline covering the firm from January 2012 to December 2023.

	(1)	(2)
	$r_{i,t+1}$	$r_{i,t+1}$
$\text{LLM}_{i,t}$	0.154***	0.0733**
	(11.28)	(2.37)
$\text{Small}_{i,t}$	-0.128	0.00502
	(-0.89)	(0.03)
$\text{LLM}_{i,t} \times \text{Small}_{i,t}$	0.263***	-0.316**
	(4.23)	(-2.03)
$\text{LAP}_{i,t}$		-0.8743**
		(-2.21)
$\text{LLM}_{i,t} \times \text{LAP}_{i,t}$		1.116***
		(2.67)
$\text{LAP}_{i,t} \times \text{Small}_{i,t}$		-1.906
		(-1.00)
$\text{LLM}_{i,t} \times \text{LAP}_{i,t} \times \text{Small}_{i,t}$		7.910***
		(3.53)
Firm FE	Yes	Yes
Date FE	Yes	Yes
R^2	0.180	0.181
N	91,361	91,361

Table IV. Regression of Next Day Returns on LLM Prediction, LAP, Conditional Probability in Response, and Confidence

This table reports regression results that assess the incremental predictive power of LAP and alternative confidence measures for next-day stock returns. The main predictors include the LLM prediction $\text{LLM}_{i,t}$, the LAP measure $\text{LAP}_{i,t}$, the probability of the first generated token $p(w_{i,N+1} \mid w_{i,\leq N})$, and the model's stated confidence $\text{Confidence}_{i,t}$. The variable $\text{LLM}_{i,t}$ takes values of -1, 0, or $+1$, indicating negative, neutral, or positive prediction, respectively. The measure $\text{LAP}_{i,t}$ is computed by averaging the model's token probabilities for the hardest-to-predict $K = 20\%$ of prompt tokens. The term $p(w_{i,N+1} \mid w_{i,\leq N})$ captures the probability of the model's first generated token, conditional on the headline (Chen et al., 2024). The variable $\text{Confidence}_{i,t}$ reflects the model's reported confidence in its prediction. Columns (1)–(4) report specifications that sequentially add these measures and their interactions to assess whether the predictive content of the LLM signal depends on LAP, the probability of the first generated token, or model-reported confidence. All regressions include firm and time fixed effects, and standard errors are clustered by date. The sample includes all U.S. common stocks with at least one news headline covering the firm from January 2012 to December 2023.

	(1)	(2)	(3)	(4)
	$r_{i,t+1}$	$r_{i,t+1}$	$r_{i,t+1}$	$r_{i,t+1}$
$\text{LLM}_{i,t}$	-0.499**	-0.702***	-3.425***	-3.491***
	(-2.21)	(-3.05)	(-11.11)	(-11.15)
$\text{LAP}_{i,t}$		-1.305***		-1.179**
		(-2.63)		(-2.41)
$\text{LLM}_{i,t} \times \text{LAP}_{i,t}$		2.860***		1.772***
		(4.85)		(3.12)
$p(w_{i,N+1} \mid w_{i,\leq N})$	0.143	0.143		
	(0.93)	(0.93)		
$\text{LLM}_{i,t} \times p(w_{i,N+1} \mid w_{i,\leq N})$	0.714***	0.710***		
	(3.13)	(3.12)		
$\text{Confidence}_{i,t}$			0.519*	0.575**
			(1.80)	(2.00)
$\text{LLM}_{i,t} \times \text{Confidence}_{i,t}$			4.520***	4.442***
			(11.46)	(11.38)
Firm FE	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes
R^2	0.179	0.180	0.183	0.184
N	91,361	91,361	91,361	91,361

Table V. Regression of Next-Period Returns on LLM Prediction Scores: In-Sample and Out-of-Sample Analysis Using Llama-2

This table reports regression results evaluating the predictive performance of LLM predictions in both in-sample (IS) and out-of-sample (OOS) periods using the Llama-2-70B model. All variables are standardized separately within the in-sample and out-of-sample periods. The in-sample period spans January 2012 to September 2022, and the out-of-sample period covers September 2023 to December 2024; definitions of these periods are provided in Table A.1. Columns (1) and (3) report in-sample estimates, while Columns (2) and (4) report the corresponding out-of-sample estimates using an evaluation window that does not overlap with the training period. Columns (3)–(4) assess whether the role of LAP in enhancing predictive performance carries over from the in-sample period to the out-of-sample period. All regressions include firm and time fixed effects, and standard errors are clustered by date.

	(1)	(2)	(3)	(4)
	$r^{\text{std}}_{i,t+1}$	$r^{\text{std}}_{i,t+1}$	$r^{\text{std}}_{i,t+1}$	$r^{\text{std}}_{i,t+1}$
$\text{LLM}^{\text{std}}_{i,t}$	0.0487***	0.0489***	0.0839***	0.0838***
	(12.09)	(12.10)	(7.45)	(7.47)
$\text{LAP}^{\text{std}}_{i,t}$		-0.00163		0.000945
		(-0.50)		(0.09)
$\text{LLM}^{\text{std}}_{i,t} \times \text{LAP}^{\text{std}}_{i,t}$		0.0120***		-0.00736
		(3.44)		(-0.74)
OOS Bootstrap p-value		[0.033]		
Sample Period	In-sample	In-sample	Out-of-sample	Out-of-sample
Firm FE	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes
Observations	82,234	82,234	9,942	9,942

Table VI. Regression of Future Capital Expenditures on LLM Predictions and LAP

This table reports regression results from the specification

$$\text{CapEx}_{i,q+2} = \alpha_i + \beta_q + \gamma \cdot \text{LLM}_{i,q} + \delta \cdot (\text{LLM}_{i,q} \times \text{LAP}_{i,q}) + \eta \cdot \text{LAP}_{i,q} + u_{i,q+2},$$

where $\text{CapEx}_{i,q+2}$ denotes firm i's capital expenditures two quarters ahead, scaled by total assets. The terms α_i and β_q represent firm and quarter fixed effects, respectively. The variable $\text{LLM}_{i,q}$ is the model-generated prediction score for firm i in quarter q, taking values of -1, -0.5, 0, 0.5, or 1, corresponding to significantly decrease, slightly decrease, no change, slightly increase, and significantly increase. The measure $\text{LAP}_{i,q}$ is computed by averaging the model's token probabilities for the hardest-to-predict $K = 20\%$ of prompt tokens. Column (1) reports results from a specification that excludes the interaction between $\text{LLM}_{i,q}$ and $\text{LAP}_{i,q}$. Column (2) adds this interaction term to assess whether the predictive content of the LLM signal varies with LAP. All regressions include firm and quarter fixed effects, and standard errors are clustered by firm.

	(1)	(2)
	$\text{CapEx}_{i,q+2}$	$\text{CapEx}_{i,q+2}$
$\text{LLM}_{i,q}$	0.798***	0.514***
	(15.89)	(5.88)
$\text{LAP}_{i,q}$		-0.0160
		(-0.57)
$\text{LLM}_{i,q} \times \text{LAP}_{i,q}$		0.148***
		(3.59)
Firm FE	Yes	Yes
Quarter FE	Yes	Yes
R^2	0.642	0.643
N	74,338	74,338

A Online Appendix

A.1 Llama Model Family

Table A.1. Timeline of the Llama model family.

The data cutoff marks the latest date of information included in the model's training data, while the release date shows when the model became publicly available. Chat versions may contain limited knowledge beyond the cutoff due to additional tuning and reinforcement learning from human feedback (RLHF).

Model	Knowledge Cutoff	Latest Update Date	In-Sample Period	Out-of-Sample Period
Llama 2 (70B)	Sept 2022	Aug 2023	Jan 2012 to Sept 2022	Sep 2023 to Dec 2024
Llama 3.3 (70B)	Dec 2023	Dec 2024	Jan 2012 to Dec 2023	–

A.2 Variable definitions

Table A.2. Variable Definitions

This table summarizes the construction and definitions of all variables used in the analysis, including LLM prediction, the LAP measure, the model's predicted capital spending, and realized future capital expenditure.

Variable	Definition
	News Headlines Predicting Stock Returns
$r_{i,t+1}$	Next-period stock return for firm i, measured in percentage points. If news is released before 4:00 p.m. on day t, the return is computed between the closing prices of day t and day $t+1$. If news is released after 4:00 p.m. on day t, the return is computed between the closing prices of day $t+1$ and day $t+2$.
$\text{LLM}_{i,t}$	Prediction score generated by the LLM for firm i at time t. Takes discrete values of -1, 0, or $+1$, indicating negative, neutral, or positive prediction, respectively.
$\text{LAP}_{i,t}$	Lookahead Propensity measure based on the model's token probabilities for the hardest-to-predict $K = 20\%$ of prompt tokens.

Continued on next page

Table A.3 – continued from previous page

Variable	Definition
$p(w_{i,N+1} \mid w_{i,\leq N})$	Probability that the model generates the first output token $w_{i,N+1}$ (e.g., "good", "bad", or "neutral") given the input prompt $w_{i,\leq N}$. It reflects the model's inner confidence (Chen et al., 2024).
$\text{Confidence}_{i,t}$	LLM's self-reported confidence in its prediction for firm i at time t, expressed as a percentage.
$\text{Small}_{i,t}$	Binary indicator equal to 1 if firm i falls into the bottom quintile of market capitalization based on trading-day $t-1$ market cap, and 0 otherwise.
	Earnings Call Predicting Capital Expenditure
$\text{CapEx}_{i,q+2}$	Capital expenditures reported at the end of the quarter $q+2$, scaled by total book assets.
$\text{LLM}_{i,q}$	Prediction generated by the LLM for firm i in quarter q based on its earnings call. The variable takes values of -1, -0.5, 0, 0.5, or 1, indicating whether future capital expenditures are expected to significantly decrease, slightly decrease, not change, slightly increase, or significantly increase before they are realized.
$\text{LAP}_{i,q}$	Lookahead Propensity measure based on the model's token probabilities for the hardest-to-predict $K = 20\%$ of prompt tokens.

B Proofs of Theoretical Results

B.1 Proof of Proposition 1

Proof. By definition,

$$\mathrm{Cov}(\tilde{y}, \widetilde{L\hat{\mu}}) = \mathbb{E}[\tilde{y} \cdot \widetilde{L\hat{\mu}}] - \mathbb{E}[\tilde{y}] \cdot \mathbb{E}[\widetilde{L\hat{\mu}}].$$

Since $\mathbb{E}[\tilde{y}] = 0$, this reduces to

$$\mathrm{Cov}(\tilde{y}, \widetilde{L\hat{\mu}}) = \mathbb{E}[\tilde{y} \cdot \widetilde{L\hat{\mu}}].$$

Using the residualized forms,

$$\tilde{y} = \varepsilon - \mathbb{E}[\varepsilon \mid \hat{\mu}, L], \quad \widetilde{L\hat{\mu}} = L^2\varepsilon - \mathbb{E}[L^2\varepsilon \mid \hat{\mu}, L].$$

Taking conditional expectation with respect to $(\hat{\mu}, L)$,

$$\begin{aligned}
\mathbb{E}[\tilde{y} \cdot \widetilde{L\hat{\mu}}] &= \mathbb{E}\Big[\mathbb{E}[\tilde{y} \cdot \widetilde{L\hat{\mu}} \mid \hat{\mu}, L]\Big] \\
&= \mathbb{E}\Big[\big(\varepsilon - \mathbb{E}[\varepsilon \mid \hat{\mu}, L]\big)\Big(L^2\varepsilon - \mathbb{E}[L^2\varepsilon \mid \hat{\mu}, L]\Big) \mid \hat{\mu}, L\Big] \\
&= L^2\,\mathbb{E}\Big[\big(\varepsilon - \mathbb{E}[\varepsilon \mid \hat{\mu}, L]\big)^2 \mid \hat{\mu}, L\Big] \\
&= L^2\,\mathrm{Var}(\varepsilon \mid \hat{\mu}, L).
\end{aligned}$$

Taking the outer expectation yields

$$\mathrm{Cov}(\tilde{y}, \widetilde{L\hat{\mu}}) = \mathbb{E}\big[L^2 \cdot \mathrm{Var}(\varepsilon \mid \hat{\mu}, L)\big],$$

which is strictly positive whenever $L > 0$. □