Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models

First Version: April 6, 2023; This Version: September 12, 2024

Abstract

We document the capability of large language models (LLMs) like ChatGPT to predict stock price movements using news headlines, even without direct financial training. ChatGPT scores significantly predict out-of-sample daily stock returns, subsuming traditional methods, and predictability is stronger among smaller stocks and following negative news. To rationalize these findings, we develop a model incorporating LLM technology, information capacity constraints, underreaction, and limits-to-arbitrage. The model establishes a critical threshold in AI capabilities necessary for profitable forecasts and predicts widespread LLM adoption can enhance market efficiency. Finally, we introduce an interpretability framework to evaluate LLMs' reasoning and accuracy.

Keywords: Large Language Models, ChatGPT, Machine Learning, Return Predictability, Textual Analysis, Market Efficiency

JEL Classification: G10, G11, G12, G14, C53

The recent proliferation of generative artificial intelligence and large language models (LLMs) like ChatGPT has been a transformative force worldwide. Although these models are primarily trained to predict the next word in a sequence, they have exhibited surprising proficiency in complex tasks, such as coding, fueling widespread interest in their emerging capabilities. However, their potential in economic applications, particularly in financial tasks, remains largely unexplored. To bridge this gap, we hypothesize and demonstrate that their acquired skills extend to the challenging task of predicting stock price movements using news headlines. Moreover, we provide theoretical results suggesting that LLMs can complement human decision-making by enhancing information processing capabilities, potentially reducing market inefficiencies, and altering information diffusion dynamics across different economic agents. Finally, we present an interpretability framework to understand how these models make correct predictions.

Our findings reveal that LLMs possess significant predictive power for economic outcomes in asset markets. For example, a self-financing daily-rebalanced strategy that buys stocks with positive ChatGPT 4 recommendations and sells stocks with negative recommendations earns a daily average return of 38 basis points (bps) pre-transaction costs, which compounds to a cumulative return of over 650% from October 2021 to December 2023.[1] Unlike prior studies that rely on supervised methods, our analysis is distinct in evaluating LLMs' ability to predict returns *without* explicit financial training.[2] Furthermore, we observe a positive relationship between LLMs' model size and economic proficiency, as LLMs with more parameters achieve higher average returns and Sharpe ratios.

The out-of-sample ability of LLMs in return prediction presents an economic puzzle. LLMs demonstrate off-the-shelf predictive power in recent periods when markets are presumed to be more efficient. To address this puzzle, we develop a theoretical model incorporating LLM technology, information processing constraints, and limits to arbitrage. Further, our framework explores the potential impact of widely available LLM forecasting capabilities

1. Assuming a transaction cost of 5 (10) basis points per round-trip trade, the strategy earns a cumulative return of over 300% (150%) over our sample period. In addition, we obtain similar results using abnormal returns from the CAPM model and the 5-factor model of Fama and French (2015) as the strategy has low loadings on risk factors.

2. See, for example, Tetlock (2007), Tetlock, Saar-Tsechansky, and Macskassy (2008), Tetlock (2011), Garcia (2013), Calomiris and Mamaysky (2019), among others.

on market dynamics by examining how these abilities might enhance investors' information processing capacity and impact overall market efficiency.

Our model generates several testable implications, all supported by our empirical findings. First, the return predictability aligns with theories of delayed information diffusion, bounded investor attention, and limits to arbitrage. For example, underreaction (and LLM predictability) is more pronounced in markets where trading is more difficult, such as with smaller stocks or short positions. Second, greater model complexity enhances information processing capabilities, so more advanced LLMs better forecast returns. Concretely, we establish the existence of a critical threshold in AI capabilities, above which these technologies can profitably predict stock returns. This profitability threshold concept is potentially applicable to various tasks and domains, such as healthcare, where AI performance directly translates into economic value. In the context of financial news, this threshold depends on news complexity, so only sophisticated LLMs should be able to exploit hard-to-understand news. Third, our model predicts an increase in how well the price reflects the underlying fundamentals once LLMs are sophisticated enough and investors start using them. Return predictability remains, however, in equilibrium, with its magnitude depending on the volatility of the non-fundamental demand and transaction costs.

Empirically, we analyze a comprehensive data set of news headlines relevant to U.S. common stocks from major news media and newswires over the period from October 2021 to December 2023. This sample period is selected to ensure an out-of-sample evaluation of LLMs like ChatGPT since the version we evaluate is trained with data up until September 2021. We prompt ChatGPT to generate a recommendation for each headline, categorizing it as positive, negative, or neutral, and use this signal to predict the following day's stock return. For instance, we form implementable strategies by entering the position at market opening and exiting at the close of the same day for news headlines released before 9 a.m. on a trading day or after 4 p.m. on the previous day. Our results demonstrate that the predictive power of ChatGPT scores is robust, extending across small and large-cap stocks, positive and negative news, and overnight and intraday news headlines. The evidence suggests that the market underreacts to company news at the time of its release, consistent with findings

in the extant literature.[3]

Notably, the predictability is more pronounced among smaller stocks and following negative news, as predicted by our model. For instance, the long and short legs of the self-financing strategy based on ChatGPT exhibit significant differences in performance, with the long leg delivering a daily average return of 9 bps and the short leg delivering 29 bps daily during our sample period.

Importantly, we find that the forecasting power of LLMs increases as the model size grows. Most basic models like GPT-1, GPT-2, and BERT display little off-the-shelf stock forecasting capabilities and no significant return predictability.[4] More advanced models, such as GPT-3.5 and DistilBart-MNLI, show some predictability but are noticeably weaker compared to the state-of-the-art model, ChatGPT 4. For example, a self-financing strategy that buys stocks with a positive ChatGPT 4 score and sells those with a negative score delivers the highest Sharpe ratio of 3.28 over our sample period, compared to 1.79 for GPT-3.5, 1.61 for DistilBart-MNLI, and negative Sharpe ratios for most basic models like GPT-1, GPT-2, and BERT. Our findings suggest that return forecasting ability is an emerging capacity of more complex LLMs, which aligns well with our theoretical prediction.

Next, we examine LLMs' forecasting capabilities across headlines with different complexities. We categorize news into low- or high-difficulty using the Flesch-Kincaid Readability Score and study how different LLMs perform with more difficult-to-understand news. Consistent with the notion that a larger number of parameters enhances LLMs' information processing capabilities, only advanced LLMs can predict returns in low-readability news. Further, we differentiate between news articles, typically independent of the firms in question, and press releases, which the firms themselves often issue. Less sophisticated models, such as BERT, demonstrate limitations in processing the latter due to their simpler algorithms, which may not adequately interpret data presented with strategic intent. Conversely, more sophisticated models like ChatGPT 4 exhibit sustained predictive accuracy, indicating their resilience to potential biases in these communications.

We then study the speed of news assimilation by analyzing the performance of the

3. For example, Bernard and Thomas (1989), Chan, Jegadeesh, and Lakonishok (1996), DellaVigna and Pollet (2009), Hirshleifer, Lim, and Teoh (2009), Jiang, Li, and Wang (2021), and Fedyk and Hodson (2023).

4. We provide details of the LLMs we examine in Appendix B of the Online Appendix.

ChatGPT-based strategies in the week after the news arrival to see how quickly stock prices react to company news. We find that ChatGPT 4 assessment scores accurately capture the immediate reaction to company news and significantly predict returns over the next two trading days but not thereafter. Thus, the market tends to underreact to the news initially and takes about two days to fully incorporate the information contained in the news.

Furthermore, considering the widespread adoption of LLMs like ChatGPT and the forecasting capabilities they demonstrate, it is imperative to assess their potential impact on market efficiency. Our theoretical model posits that LLMs can increase investors' information processing capacity and reduce market inefficiencies. As a result, the return predictability should weaken as LLMs' model size increases and more investors start using them. While testing this hypothesis represents an empirical challenge, we find some suggestive evidence: a general decline in the performance of the ChatGPT-based strategy during our sample period over which GPT models' capabilities and adoption skyrocketed. For example, its annualized Sharpe ratio drops from 6.54 in 2021Q4 to 3.68 in 2022, and to 2.33 in 2023.

Finally, we propose a novel interpretability technique that combines surrogate modeling with topic modeling to gain insights into the reasoning behind LLMs' predictions. This two-step approach can analyze both LLMs' predictions and their explanations. The first step involves surrogate modeling, a machine learning technique that employs an interpretable model, such as linear regression, to understand a more complex one (an LLM). We use linear regression for our surrogate modeling to separately analyze (i) LLM scores to identify the factors influencing their predictions, (ii) prediction performance (score multiplied by return) to uncover the factors contributing to successful or failed predictions, and (iii) differences in scores or performance across various LLMs to investigate the factors driving model improvements.

The second step involves using topic modeling to enhance the interpretability of our surrogate models by creating discrete topics that serve as dummy variables in the regression analysis. We apply topic modeling to both news headlines and LLMs' explanations. When applied to the former, the topic model reveals the underlying themes that influence LLMs' predictions; when applied to the latter, it uncovers patterns in LLMs' reasoning process.

Our interpretability analysis reveals several insights into ChatGPT's decision-making

process when predicting stock returns. News related to insider transactions and share repurchase declarations significantly impacts the model's predictions. For instance, GPT-4 tends to rate executive stock transactions overly negatively, while it accurately assesses director and chairperson transactions, leading to outperformance. Share repurchase announcements are correctly identified by GPT-4 as positive signals, resulting in substantial performance gains. However, the model struggles in some areas, for instance, in interpreting the impact of convertible note offerings, where it often overestimates their positive effects. Interestingly, GPT-4 shows marked improvements over GPT-3.5, both in general and in specific areas such as interpreting reverse stock splits and industry-specific news in sectors like electric vehicles. These findings not only shed light on the strengths and limitations of LLMs in financial prediction tasks but also highlight the evolving capabilities across different LLMs. Overall, we believe this technique can be applied to any LLM-related task involving text inputs or explanations and should help better understand these "black boxes."

Evidence of LLMs' forecasting capabilities provides several economic insights. First, their effectiveness hints at potential information processing advantages approaching or surpassing human analysts with labor market implications that we leave for future research. Second, the success of LLMs in this domain underscores the emergent reasoning capabilities of state-of-the-art AI systems in economic tasks and shows how AI technologies may alter information diffusion and decision-making processes in various economic contexts. Third, our study could benefit many investors by providing empirical evidence on the efficacy of LLMs in predicting stock market returns. Fourth, regulators should be aware of LLMs' potential effects on market behavior, information dissemination, and price formation. Finally, understanding LLMs' ability to forecast returns sheds light on market efficiency and limits to arbitrage.

Related Literature

The application of LLMs in economics, particularly ChatGPT, is a relatively unexplored area. In addition to our study, recent research on ChatGPT in economics includes Korinek (2023), Hansen and Kazinnik (2023), Bybee (2023), Noy and Zhang (2023), and Manning,

Zhu, and Horton (2024), each addressing different questions than ours.[5] Furthermore, a contemporaneous complementary work by Chen, Kelly, and Xiu (2023) employs a supervised two-step procedure by first embedding news articles in a high-dimensional space using different embedding techniques, including LLM-based embeddings, and then using these embeddings as inputs for a forecasting model to predict stock returns. While effective, their method requires sophisticated implementation and is better suited for skilled investors. In contrast, we focus on testing the off-the-shelf capabilities of LLMs (accessible to most investors) and providing a theoretical model to explain their potential economic impact on financial markets. Both approaches are complementary: the direct application of conversational AI systems offers accessibility and interpretability, while the two-step procedure provides flexibility in customization and optimization for specific prediction tasks.

We also contribute to the literature that employs machine learning to study finance research questions, including textual analyses of news articles to extract sentiment and predict stock returns.[6] Our unique contribution is being the first to provide a theoretical framework on LLMs and their potential impact on market dynamics and document comprehensive empirical evidence on LLMs' stock return predictability skills. Our model adds to the literature on information incorporation into market prices and introduces LLM technology.[7] Our paper also relates to the literature on employment exposures and vulnerability to AI-related technology by documenting an important task in the financial industry where off-the-shelf LLMs perform well.[8] Finally, our results, along with the interpretability technique we propose

5. Moreover, Xie et al. (2023) find ChatGPT is no better than simple methods such as linear regression when using numerical data in prediction tasks, and Ko and Lee (2023) try to use ChatGPT to help with a portfolio selection problem but find no positive performance. Both results are unsurprising since ChatGPT is better at text-based tasks. In addition, Eisfeldt, Schubert, and Zhang (2023) study the impact of ChatGPT on firm value through the channel of labor productivity. Finally, Kogan et al. (2023) utilize ChatGPT to classify job tasks into routine vs. non-routine categories, and W. Jiang et al. (2023) use ChatGPT to evaluate the complementary vs. substitutive impact of fintech innovations on different occupations.

6. See, e.g., Jegadeesh and Wu (2013), Rapach, Strauss, and Zhou (2013), Hoberg and Phillips (2016), Baker, Bloom, and Davis (2016), Manela and Moreira (2017), Hansen, McMahon, and Prat (2018), Gu, Kelly, and Xiu (2020), Ke, Kelly, and Xiu (2019), Ke, Montiel Olea, and Nesbit (2019), F. Jiang et al. (2019), Cohen, Malloy, and Nguyen (2020), Freyberger, Neuhierl, and Weber (2020), Bybee et al. (2023), Lopez-Lira (2023), and Cong, Liang, and Zhang (2024).

7. For example, Van Nieuwerburgh and Veldkamp (2010), Kyle (1985, 1989), Verrecchia (1982), Dávila and Parlatore (2018), and Dávila and Parlatore (2021).

8. Recent works by Agrawal, Gans, and Goldfarb (2019), Webb (2019), Acemoglu et al. (2022), Acemoglu and Restrepo (2022), Acemoglu (2024), Babina et al. (2024), W. Jiang et al. (2023), Kogan et al. (2023), and Noy and Zhang (2023) have examined the extent of job exposure and vulnerability to AI-related technology

to evaluate LLMs' reasoning, contribute to understanding LLMs' potential when predicting stock market returns, which can inspire future research on developing LLMs tailored to the financial industry's needs.[9]

1 Institutional Background

ChatGPT is a large-scale language model developed by OpenAI based on the GPT (Generative Pre-trained Transformer) architecture. It is one of the most advanced natural language processing (NLP) models developed so far and was trained on a massive corpus of text data to understand the structure and patterns of natural language. The Generative Pre-trained Transformer (GPT) architecture is a deep learning algorithm for natural language processing tasks. It was developed by OpenAI and is based on the Transformer architecture, which was introduced in Vaswani et al. (2017). The GPT architecture has achieved state-of-the-art performance in various natural language processing tasks, including language translation, text summarization, question answering, and text completion.

The GPT architecture uses a multi-layer neural network to model the structure and patterns of natural language. It is pre-trained on a large corpus of text data, such as Wikipedia articles or web pages, using unsupervised learning methods. This pre-training process allows the model to develop a deep understanding of language syntax and semantics, which is then fine-tuned for specific language tasks. One of the unique features of the GPT architecture is its use of the transformer block, which enables the model to handle long sequences of text by using self-attention mechanisms to focus on the most relevant parts of the input. This attention mechanism allows the model to understand the input context better and generate more accurate and coherent responses.

ChatGPT has been trained to perform various language tasks such as translation, summarization, question answering, and generating coherent and human-like text. ChatGPT's ability to generate human-like responses has made it a powerful tool for creating chatbots and virtual assistants to converse with users. While ChatGPT is a powerful tool for general-

as well as the consequences for employment and productivity.

9. See, for example, Wu et al. (2023) on a large language model for finance—BloombergGPT. In addition, see Lerner et al. (2024) for a recent study on increasingly valuable financial innovations in the U.S.

purpose language-based tasks, it is not explicitly trained to predict stock returns.

In addition to evaluating ChatGPT, we also assess the capabilities of other prominent natural language processing models such as BERT, BART, DistilBart-MNLI, and FinBERT. By considering the more basic models alongside ChatGPT, we can examine the importance of model complexity for stock return prediction based on textual data. Appendix B of the Online Appendix provides an overview of the LLMs we study in this paper.

2 Model

We use an economic model to explore the return predictability implications of using LLMs when analyzing news, building on the multi-period noisy rational expectations framework of Grundy and McNichols (1989) and Brown and Jennings (1989) and in a similar setup as Gromb and Vayanos (2010). Our model extends their work by introducing LLMs as novel information processors and characterizing the equilibrium in their presence.

We establish the existence of a threshold model size above which LLMs can profitably predict returns. Furthermore, we show that the price better reflects the fundamentals as the proportion of agents using LLMs and LLMs' model size increase, highlighting the LLMs' potential to enhance market efficiency and reduce return predictability. All proofs are in Appendix D of the Online Appendix.

2.1 Agents

We have two types of agents, attentive and inattentive, indexed by A and I, respectively, an asset with uncertain dividends and news revealed at the beginning of each period. Both agents have CARA utility with the same risk aversion (α) and are price-takers. The asset is in zero net supply so that we can focus on expectations rather than risk premiums. Incorporating two types of agents in the model allows us to examine how the adoption of LLMs by different market participants influences return predictability and overall market efficiency.

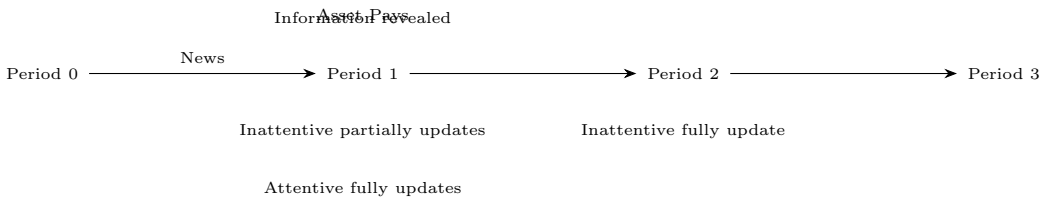
The total measure of agents paying attention to the asset is denoted by $V > 0$. V closely resembles the potential volume of traders and should be lower in smaller stocks and, by definition, in illiquid stocks. Attentive agents are a fraction $\pi_A \in (0, 1)$ of this population

and have a total measure of $\pi_A V$. Inattentive agents a fraction of $\pi_I = 1 - \pi_A$ and a total population of $\pi_I V$. We model V as fixed for a given market so that π_A can be determined in equilibrium based on the costs of participating as an attentive investor and the attractiveness of speculation.

2.2 Time Dynamics

The model features three periods that model short trading horizons. In period one, unexpected news about the company is released. Attentive agents incorporate this information in their forecasts better than inattentive investors. We assume inattentive agents only update their forecasts partially (because of information capacity constraints) and do not consider the price when updating since they have just enough information capacity to understand a fraction of the news implications. In period two, both agents update their expectations entirely. Finally, in period three, the asset pays off a dividend, $\tilde{d}$, and the economy ends. There is no intertemporal discount for notational ease since we are modeling very short horizons.

Attentive agents know that inattentive agents do not update entirely in the first period but will do so in the second. Hence, they can trade in the first period to take advantage of the price discrepancy. Without an additional source of uncertainty, this would be a riskless trade, which is unrealistic. We assume that when trading with a short horizon, nonfundamental trades may move the price further away from its fundamental value (e.g., De Long et al. (1990)). That is, there is nonfundamental risk stemming from noise traders, normally distributed with an expected value of zero and variance of σ_u^2.[10]



10. Note that we only add noise traders in the last period for tractability.

2.3 Equilibrium Before News

The asset payoff is random and is given by

$$\tilde{d} = \mu_d + \sigma_\xi \tilde{\xi},\ \tilde{\xi} \sim N(0,1). \tag{1}$$

We refer to the expected value of the asset payoff, μ_d, as the asset fundamentals. There are two sources of uncertainty. First, agents do not perfectly know the fundamentals, μ_d. Second, there is real variation in the asset's payoff given by $\sigma_\xi \tilde{\xi}$. We assume independence between the uncertainty about the fundamentals and the real variation. While agents do not know the fundamentals, they have a correct prior. The prior is correct in the sense that it is equal to the population distribution:

$$\mu_d \sim N(\bar{d}, \sigma_d). \tag{2}$$

We let $x_{j,t+1}$ denote the gross demand of the asset for each type of investor in period t. Without unexpected news, the demand of both agents is given by:

$$x_{1,j} = \frac{\bar{d} - p_0}{\alpha(\sigma_d^2 + \sigma_\xi^2)}. \tag{3}$$

The asset is in zero net supply and we obtain the equilibrium price by the market clearing condition:

$$V \frac{\bar{d} - p_0}{\alpha(\sigma_d^2 + \sigma_\xi^2)} = 0 \tag{4}$$

Hence, the price at period zero is given by its unconditional expected value:

$$p_0 = \bar{d}. \tag{5}$$

2.4 Information Structure

Consider a scenario where news about a company gets released unexpectedly. Agents interpret this news as a signal with information about the fundamentals. The amount of

information they use is related to their information processing capacity and the news complexity.

We assume the news' maximum information content about the fundamental contains a total precision of τ_S.

$$s = \mu_d + \varepsilon, \quad \varepsilon \sim N(0, \sigma_S^2 = \frac{1}{\tau_S}) \tag{6}$$

All agents, including attentive, inattentive, and LLMs, observe the same signal realization s, representing the company's news.[11] However, they process this signal differently due to their varying information processing capacities. We implicitly assume agents are not biased when processing the information by assuming everyone observes the same signal. We could incorporate bias at the cost of additional notation.[12]

The precisions for each type of agent for the new fundamental are given by:

$$\tau_A = \gamma_A \tau_S, \ \tau_I = \omega \tau_A, \ \tau_L = \lambda(c,k)\tau_S, \tag{7}$$

where $\gamma_A \in (0,1)$ is attentive agents' information capacity, and $\omega \in (0,1)$, describes how good inattentive agents' capacity is relative to attentive agents. $\lambda(c,k) : \mathbb{R}_2^+ \to (0,1]$ is a twice differentiable, decreasing function of news complexity c and an increasing function of model size k, and represents LLMs' information capacity.

We additionally assume $\frac{\partial \lambda}{\partial c} \leq 0$, $\frac{\partial \lambda}{\partial k} > 0$, $\frac{\partial^2 \lambda}{\partial c^2} < 0$, $\frac{\partial^2 \lambda}{\partial k^2} < 0$, $\frac{\partial^2 \lambda}{\partial c \partial k} > 0$, $\lambda(c,0) = 0 \ \forall \ c \geq 0$, and $\lim_{k\to\infty} \lambda(c,k) = 1 \ \forall \ c \geq 0$. As news complexity increases, the information processing ability of LLMs decreases. However, as the model size of LLMs increases, their information processing capacity improves, even in the presence of complex news. When there is no confusion given the context, we use λ to denote $\lambda(c,k)$.

Attentive agents process the signal with precision τ_A in the first period. In contrast, inattentive agents process the signal with precision $\omega\tau_A$ in the first period and the remaining precision $(1-\omega)\tau_A$ in the second period. That is, inattentive agents only process a fraction of the information of attentive agents given by ω in the first period and the remaining fraction

11. Hence, our work differs from the literature studying how private information gets incorporated into prices (e.g., Kyle (1985) and Shleifer and Vishny (1997)).

12. For example, different signal components for the signal of inattentive agents at the first and second period and the attentive agents' signal ($s_{I,1}$, $s_{I,2}$, and s_A) would allow for reversal return dynamics at the cost of a more complex model structure.

in the second. Intuitively, inattentive investors see some news, have an initial reaction, and trade on it. After some time, say a few hours later, these inattentive investors will better understand the news' full implications and can trade again. The signal inattentive investors processed is entirely contained in the signal the attentive investors processed.

Attentive agents only process a fraction γ_A of the total information. When $\gamma_A < 1$, even attentive agents cannot process the complete information, resulting in return predictability from the perspective of an agent who can process the entire signal. We analyze this case later by incorporating a hypothetical LLM with a large enough model size, such that $\lambda(c, k) > \gamma_A$.

LLMs process the signal with precision τ_L in the first period. We assume that if $\tau_L < \tau_A$, then the rest of the precision $\tau_A - \tau_L$ will be processed in the second period (intuitively, one could think that a human would revise the output); otherwise, there is just one update for LLMs in the first period.

We assume that inattentive investors do not condition their forecasts on prices. Attentive investors have the complete information set available in the economy, so they do not need to learn from the price unless there are LLMs with information capacity greater than theirs and attentive agents are not using LLMs.

2.5 Information Update

The updated expectation for each agent in period one is then given by

$$\mu_{j|s} \equiv E[d|s_j] = \frac{\bar{d}\tau_d + s\tau_j}{\tau_d + \tau_j}, j \in \{A, I, LLM\}, \tag{8}$$

and the updated fundamental variance is given by

$$Var[\mu|s_j] = \sigma^2_{\mu,j|s} \equiv \frac{1}{\tau_{\mu,j|s}} \equiv \frac{1}{\tau_d + \tau_j},\ j \in \{A, I, LLM\}. \tag{9}$$

Moreover, the updated total variance is

$$\sigma^2_{d,j|s} \equiv Var[d|s_j] = \sigma^2_{\mu,j|s} + \sigma^2_{\xi}. \tag{10}$$

2.6 LLMs for Forecasting

We first focus on the case where agents are not using LLMs to trade and are only available to the econometrician. Intuitively, we are modeling the case of how LLMs react (out-of-sample) to the information, closely aligning with the post-September 2021 period, the knowledge cutoff date for ChatGPT-4, but before March 2023, when ChatGPT-4 was released. We first characterize the price and return dynamics and then explore how LLMs forecast. In a later subsection, we make LLMs available to agents and analyze the resulting dynamics.

Attentive agents understand there is an initial underreaction. Hence, in the first period, they will trade to take advantage of the difference between the future and current prices. In the second period, they will trade considering the dividend. Because of the dynamic nature of the problem, it is easier to solve the model starting from the second period.

2.6.1 Second Period

In the second period, there are no further signals for the attentive agent, and the inattentive agent fully updates, so both agents have the same fundamental expectations given by

$$\mu_{A|s} \equiv E[\tilde{d}|s_A] = \frac{\bar{d}\tau_d + s\tau_A}{\tau_d + \tau_A}. \tag{11}$$

Both types of agents also believe the fundamental has variance given by:

$$\sigma^2_{\mu,A|s} \equiv \frac{1}{\tau_{A|s}} \equiv Var[d|s_A] = \frac{1}{\tau_d + \tau_A}. \tag{12}$$

With CARA utility, their portfolio problem does not depend on their wealth level and is independent of previous period prices. Their (gross) demand for the asset is given by:

$$x_{3,j} = x_3 \equiv \frac{\mu_{A|s} - p_2}{\alpha\sigma^2_{d,A|s}} \; j \in \{A, I\}. \tag{13}$$

Hence, the price is given by

$$p_2 = \mu_{A|s} + u_2\frac{\alpha\sigma^2_{d,A|s}}{V}. \tag{14}$$

Moreover, the expected value of the price for attentive agents as of period 1 is

$$E_{1,A}[p_2] = E_{1,A}[\mu_{A|s} + u_2 \frac{\alpha \sigma^2_{d,A|s}}{V}] = \mu_{A|s} = E[\tilde{d}|s_A]. \tag{15}$$

The expectation of time two is measurable at time one since there are no further updates on attentive agents. Hence, the only source of risk in the price is the nonfundamental demand, which is normally distributed with variance and precision given by:

$$\sigma_p^2 \equiv Var_{1,A}[p_2] = \alpha^2 \frac{\sigma_u^2}{V^2} (\sigma^2_{d,A|s})^2, \tag{16}$$

$$\tau_p \equiv \frac{1}{\sigma_p^2} = \frac{1}{Var_{1,A}[p_2]} = \frac{V^2 \tau_u}{\alpha^2 (\sigma^2_{d,A|s})^2}. \tag{17}$$

Thus, the expected price is equal to the expected dividend. Also, the nonfundamental risk is amplified by the fundamental risk of the asset and the risk aversion and diminished by the total volume of traders.

2.6.2 Period One

In period 1, attentive agents understand the news and trade to profit from the differences in future and current prices. In contrast, inattentive agents will trade to get the final period payoff, not realizing they do not fully understand the impact of news until the second period.

Attentive agents Let $\tilde{w_3} = w_2 + x(\tilde{d} - p_2)$ and $\tilde{w_2} = w_1 + x_2(\tilde{p_2} - p_1)$. Given the demand in period 2 and the equilibrium prices, the value function of period 2 for attentive agents is:

$$V2(w_2) = \max_x E[U(\tilde{w_3})] = \max_{x_3} -exp\{-\alpha E[\tilde{w_3}] + \frac{\alpha^2}{2} Var(\tilde{w_3})\} \tag{18}$$

In period 1, attentive agents maximize the expected period-2 value function:

$$V1(w_1) = \max_{x_2} E_1[V2(\tilde{w_2})] \tag{19}$$

In the Online Appendix, we show that we can write the optimal demand function as

$$x_{2,A} = \frac{(\mu_{A|s} - p_1)}{\alpha(\sigma^2_{d,A|s})} + \frac{(\mu_p - p_1)}{\alpha\sigma^2_p}. \tag{20}$$

The first part corresponds to the traditional demand from CARA investors for an asset that pays a dividend, and the second part is due to the opportunity to take advantage of temporary mispricing. The first part arises because, given the second period's desired demand, there is an opportunity to obtain a potentially lower price and hold the asset until the dividend is realized.

Another interpretation is viewing agents as more confident regarding their signals. Let $\tau_p = \frac{1}{\sigma^2_p}$ and $\tau_{d,A|s} = \frac{1}{\sigma^2_{d,A|s}}$. We can rewrite the demand as

$$x_{2,A} = (\tau_{d,A|s} + \tau_p)\frac{\mu_{A|s} - p_1}{\alpha} \equiv \tau_{p,d}\frac{\mu_{A|s} - p_1}{\alpha}, \tag{21}$$

where $\tau_{p,d} = \tau_{d,A|s} + \tau_p$ is the total precision.

Inattentive agents Inattentive agents' demand is given by

$$x_{2,I} = \frac{(\mu_{I|s} - p_1)}{\alpha(\sigma^2_{I|s} + \sigma^2_\xi)} = \tau_{d,I|s}\frac{(\mu_{I|s} - p_1)}{\alpha}. \tag{22}$$

Equilibrium Price If we define weighted precisions as $\tau_{\pi_A} = \pi_A\tau_{p,d}$, $\tau_{\pi_I} = \pi_I\tau_{d,I|s}$, then we can express the the first-period equilibrium price as

$$p_1 = \frac{\mu_{A|s}\pi_A\tau_{p,d} + \mu_{I|s}\pi_I\tau_{d,I|S}}{\pi_A\tau_{p,d} + \pi_I\tau_{d,I|S}} = \frac{\mu_{A|s}\tau_{\pi_A} + \mu_{I|s}\tau_{\pi_I}}{\tau_{\pi_A} + \tau_{\pi_I}} \tag{23}$$

$$\equiv \mu_E, \tag{24}$$

where $\mu_E = \frac{\mu_{A|s}\tau_{\pi_A} + \mu_{I|s}\tau_{\pi_I}}{\tau_{\pi_A} + \tau_{\pi_I}}$ is the economy-wide expectation. The price is the dividend expectation weighted by each type of agent's (weighted) precisions.

To determine the equilibrium proportion of informed agents, we can assume, as in Grossman and Stiglitz (1980), that there is a fixed cost of participating in this market as an attentive investor of $c_A \geq 0$, and this cost is higher when the investor wants to short. π_A

adjusts until attentive agents are indifferent between being attentive and paying c_A or not participating in the market. We model c_A as an exogenous constant, which is higher when participating as an attentive agent that can short stocks. These costs can stem from trading fees and investment restrictions, as only certain participants can short stocks.[13]

Period 1 Unconditional Returns The immediate (unconditional) expected jump from the non-news price to the news price is

$$E[r_1] = E[p_1] - E[p_0] = E[\mu_E] - \bar{d} \tag{25}$$

$$= \bar{d} - \bar{d} = 0, \tag{26}$$

which corresponds to the returns we would observe in high-frequency data. These returns are unconditionally 0 as the agents are not biased regarding the signal. For this simple economy, it means that the econometrician needs to condition a signal to study the properties of period 1 returns.[14]

Period 1 to Period 2 Conditional Returns The expected price of period 2 is simply $\mu_{A|s}$. Hence, the expected (dollar) returns, conditional on period 1's information set (not necessarily observable to the econometrician), between period 2 and period 1 are given by:

$$E_1[r_2] = E_1[p_2 - p_1] = \mu_{A|s} - \mu_E. \tag{27}$$

The unconditional period 1 to period 2 return is zero on average since the signals are unbiased. That is, without taking a stance on whether the news is good or bad, it is not possible to take a profitable strategy.

Market Efficiency We define pricing errors, α_M, as the difference between the expected value of the dividend conditional on the signal, and the price at period 1.

13. The equilibrium exists because attentive agents' expected utility (and profits) are decreasing in the proportion of attentive agents. If c_A is too high, there may not be any attentive investors. There is no paradox in this case since we assume news is produced exogenously within the short horizon.

14. For simplicity, we assume there are no nonfundamental traders in the first period, which implies there is always momentum due to the delay in attention within the model, and this assumption can be relaxed at the cost of more algebra.

$$\alpha_M \equiv E[\tilde{d}|s] - E[p_1|s]. \tag{28}$$

And we define mispricing as the expected square value of the pricing errors, where the expectation is taken with respect to the signal s:

$$E[\alpha_M^2] \equiv E[(E[\tilde{d}|s] - E[p_1|s])^2]. \tag{29}$$

Proposition 2.1. *Mispricing is*

1. *Decreasing in the proportion of attentive agents, the information capacity of either type of agent, and the total volume of traders.*

2. *Increasing in the risk aversion and the noise trader risk.*

Mispricing is lower when there are more attentive trades, either because of more attentive agents or lower risks to trade, such as when the volume of traders is higher, causing the impact of noise traders to be smaller. Moreover, we define markets as more efficient if there is less mispricing. However, we cannot measure mispricing directly because we do not have access to the subjective information set. Fortunately, LLMs provide an imperfect proxy for this information set.

2.6.3 Predictability with LLMs

Using an LLM, the expectation about the dividend is given by

$$\mu_L \equiv E_L[d|s] = \frac{\bar{d}\tau_d + s\lambda\tau_S}{\tau_d + \lambda\tau_S}, \tag{30}$$

and the updated precision is $\tau_{\mu,L|s} = \tau_d + \lambda\tau_S$.

From the perspective of a CARA agent using LLMs to predict the fundamental, the optimal demand is given by

$$x_{2,L} = \tau_{d,L|s}\frac{(\mu_{L|s} - p_1)}{\alpha}, \tag{31}$$

with an expected subjective profit (without price impact and conditional on period one prices) of

$$x_{2,L}(\mu_{L|s} - p_1) = \tau_{d,L|s}\frac{(\mu_{L|s} - p_1)^2}{\alpha}. \tag{32}$$

However, this strategy has an actual profit (without price impact) of

$$x_{2,L}(\mu_A - p_1) = \tau_{d,L|s}\frac{(\mu_{L|s} - p_1)(\mu_A - p_1)}{\alpha}, \tag{33}$$

with the unconditional expectation (over the signal) given by

$$E[x_{2,L}(\mu_A - p_1)] = E[\tau_{d,L|s}\frac{(\mu_{L|s} - p_1)(\mu_A - p_1)}{\alpha}]. \tag{34}$$

Conditional on the signal, the strategy is profitable when $(\mu_{L|s} - p_1)$ has the same sign as $(\mu_A - p_1)$, which happens when the LLM correctly interprets the news as good or bad. Equation 33 closely maps to the strategy we implement in the empirical strategy in Section 5 of the paper. Theorem 1 characterizes its profitability properties.[15] We say that there is return predictability if the profits are positive, and it is increasing if the profits are increasing.

Theorem 1. *In the case when only the econometrician has access to LLMs, the profitability of LLM-based strategies is increasing in the LLM's model size. Moreover, for a fixed set of parameters and news complexity, there is a unique threshold of model size, k^*, such that only larger LLMs with $k > k*$ can predict returns profitably. The threshold k^* is:*

1. *Increasing in inattentive agents information capacity ω, attentive agents information capacity γ_A, the proportion of attentive agents π_A, and the total volume of agents V.*

2. *Decreasing in agents' risk aversion, α, and the noise trader variance σ_u^2.*

Conditional on the strategy being profitable, the profitability is decreasing in inattentive agents' information capacity ω, the proportion of attentive agents π_A, and increasing in the noise trader variance σ_u^2 and agents' risk aversion. Hence, we have the following propositions.

15. If we add noise traders in the first period, the profitability is still increasing in model size, but for sufficiently small volume, even inattentive agents can make profits on average since the price is generally far away from the fundamental value, so even having the correct prior and no signal is enough.

Proposition 2.2. *If LLM-return predictability is profitable, LLM-return predictability is larger in markets with a lower proportion of attentive market participants, such as smaller or illiquid stocks, and with negative news.*

Proposition 2.3. *Holding model size and news complexity constant, and if the LLM-based strategy is profitable, a decline in its returns due to better information processing of inattentive agents, higher volume of traders, a higher proportion of attentive market participants, or lower noise traders' risk corresponds to a reduction in mispricing.*

These propositions help us know when using LLM return predictability should be lower and when we can use this decline as a proxy for decreased mispricing.

To get the intuition of some of the results, Figure 1 presents how the expected profitability of LLM-based strategies differ when changing one of the following parameters: the information capacity of the LLM technology (λ) or the information capacity of the inattentive agents (ω).[16] The plots show that LLM-return predictability is monotonically increasing in LLM's information capacity and decreasing in inattentive agents' information capacity if the predictability is positive.

[Insert Figure 1 about here]

2.7 LLMs Available for Trading

The previous section assumes only the econometrician has access to LLMs and hence abstracts from the price impact associated with trading based on LLMs' signals. Now we model the case where LLMs are better than inattentive investors but not better than attentive investors, $\gamma_I < \lambda(c,k) \leq \gamma_A$. Implicitly, we assume the technology is widely available, as is the case for the current best LLMs.

2.7.1 LLMs Better Than Inattentive Agents

Let's assume a fraction of inattentive agents, θ, start processing the information with LLMs, and the remaining information will be processed in the second period. Then the new price in the first period, $p_{1|L,I}$ is given by

16. We change directly the parameter λ instead of k or c for illustration purposes.

$$p_{1|L,I} = \frac{\mu_{A|s}\pi_A\tau_{p,d} + (1-\theta)\mu_{I|s}\pi_I\tau_{d,I|S} + \theta\pi_I\mu_{L|s}\tau_{d,L|s}}{\pi_A\tau_{p,d} + (1-\theta)\pi_I\tau_{d,I|S} + \theta\pi_I\tau_{d,L|s}}. \tag{35}$$

In this case, mispricing will decline as markets become more efficient. Moreover, all things equal, mispricing will decline more with more advanced LLMs.

Theorem 2. *If LLMs are more capable than inattentive traders, mispricing decreases in the proportion of inattentive agents using LLMs, and on LLMs' model size k, and hence, markets are more efficient than in the equilibrium without LLMs if inattentive agents use them.*

Intuitively, agents using LLMs that are better than them at correctly inferring the news' impact will have better expectations. The more agents use these advanced LLMs, the more the price will reflect the fundamental information.

Proposition 2.4. *If LLMs have a sufficiently large model size and all inattentive agents use them, there will no longer be any return predictability.*

When all inattentive agents use LLMs, and LLMs have the same capacity as the attentive agents, return predictability completely disappears due to the price impact.

Similar to the case where LLMs are only available to the econometrician, for a sufficient size of LLMs, there is positive return predictability. In contrast to before, however, this return predictability is not typically monotonic in the size of the LLM. It is also not monotonic in the proportion of inattentive agents using LLMs.

2.7.2 LLMs Better than Attentive Agents

We now focus on the case where LLMs become better than attentive agents at processing information, and all attentive agents switch to the better LLM technology. This case is analogous to increasing attentive agents' information capacity, γ_A.

Proposition 2.5. *If attentive agents use better-than-human LLMs, market efficiency improves as LLM model size increases.*

Intuitively, attentive agents are substituting their information processing capacity with that of the LLM's and mispricing is lower if attentive agents have a higher capacity.

An additional convoluted case occurs when only a fraction of attentive agents use LLMs. The model becomes more complex because attentive agents that do not use LLMs can learn from the price. Intuitively, institutional investors who do not use LLMs but observe prices can react to sudden price changes. The equilibrium can be solved by assuming the price is linear in the LLMs' signal. In that case, attentive agents without LLM access receive a second signal about the fundamentals, and their expectations approach the expectations of attentive agents that use LLMs. This case is interesting when considering an equilibrium with autonomous LLM trading, and the implications are similar to the above cases.

3 Data

We utilize four primary datasets for our analysis: the Center for Research in Security Prices (CRSP) daily returns, news headlines, RavenPack news database, and the NYSE Trade and Quote (TAQ) database. The sample period begins in October 2021 and ends in December 2023. This sample period ensures that our evaluation is out-of-sample as ChatGPT's training data stops in September 2021.

We obtain daily stock returns, open prices, and close prices from the CRSP. Our sample consists of all the stocks listed on the New York Stock Exchange (NYSE), the National Association of Securities Dealers Automated Quotations (NASDAQ), and the American Stock Exchange (AMEX), with at least one news story covered by a major news media or newswire. Following prior studies, we focus our analysis on common stocks with a share code of 10 or 11. Our intraday price and return data come from the TAQ database. We clean the TAQ data and construct minute-by-minute observations of intraday volume-weighted average prices following the cleaning procedures in the literature (e.g., Bollerslev, Li, and Todorov (2016)).

We first collect a comprehensive news dataset for all CRSP companies using web scraping based on the company name or the ticker. The resulting dataset comprises news headlines from various sources, such as major news agencies and financial news websites. For each

company, we collect all the news during the sample period and then match the headlines with the identifying information from RavenPack.

After the matching procedures, our final sample has 134,129 headlines of 4,106 unique companies from October 2021 to December 2023. In the sample, about 68% (or 91,700) of the headlines correspond to press releases and the remaining 32% are news articles. In addition, based on the release time, we group news into overnight vs intraday news because the entry times for the two types of news are different when forming trading strategies. About 81% (or 109,206) of the headlines are classified as overnight news as they are released either before 9 a.m. or after 4 p.m. on a trading day, and the remaining 19% are classified as intraday news.[17]

Matching with RavenPack assures that only relevant news will be used for the experiment. They closely monitor the major financial news distribution outlets and have a quality procedure matching news, timestamps, and entity names, which solves any errors that arise from the web scraping procedure. We employ the "relevance score" (between 0 and 100) from Ravenpack to indicate how closely the news pertains to a specific company. A 0 (100) score implies that the entity is mentioned passively (predominantly). Our sample requires news stories with a relevance score of 100. We limit it to complete articles and press releases and exclude headlines categorized as 'stock-gain' and 'stock-loss' as they only indicate the daily stock movement direction. To avoid repeated news, we require the "event similarity days" to exceed 90, which ensures that only new information about a company is captured.[18]

4 ChatGPT Prompt

In this section, we discuss how we prompt ChatGPT to extract information from news headlines and illustrate it with an example.

A prompt is a short text that provides context and instructions for ChatGPT to generate

17. We use 9 a.m. as the cutoff in the morning to classy overnight news so that there is time to process and trade when the market opens. Results are similar if we use alternative cutoffs such as 9:15 or 9:30 a.m.

18. Furthermore, we eliminate duplicates and overly similar headlines for the same company on the same day. We gauge headline similarity using the Optimal String Alignment metric (the Restricted Damerau-Levenshtein distance) and remove subsequent headlines with a similarity greater than 0.6 for the same company on the same day.

a response. The prompt can be as simple as a single sentence or as complex as a paragraph or more, depending on the nature of the task. Prompts enable ChatGPT to perform a wide range of language tasks, such as language translation, text summarization, question answering, and even generating coherent and human-like text. They allow the model to adapt to specific contexts, generate responses tailored to the user's needs, and perform tasks in different domains.

We use the following prompt in our study and apply it to the publicly available headlines.

> *Forget all your previous instructions. Pretend you are a financial expert. You are a financial expert with stock recommendation experience. Answer "YES" if good news, "NO" if bad news, or "UNKNOWN" if uncertain in the first line. Then elaborate with one short and concise sentence on the next line. Is this headline good or bad for the stock price of _company_name_ in the short term?*
> *Headline: _headline_*

We ask ChatGPT to assume the role of a financial expert with experience in stock recommendations. The terms _company_name_ and _headline_ are substituted by the firm name and the respective headline during the query. The prompt is specifically designed for financial analysis and asks ChatGPT to evaluate a given news headline and its potential impact on a company's stock price in the short term. ChatGPT is requested to answer "YES" if the news is good for the stock price, "NO" if it is bad, or "UNKNOWN" if it is uncertain. ChatGPT is then asked to explain in one sentence to support its answer concisely. The prompt specifies that the news headline is the only source of information provided to ChatGPT. It is not assumed that the headline contains sufficient information to reasonably assess its impact on the stock price since the model can answer it does not know. We set the temperature of GPT models to 0 to maximize the reproducibility of the results.[19]

For example, consider the following headline about Humana in December 2023:

> *Cigna Calls Off Humana Pursuit, Plans Big Stock Buyback*

19. Temperature is a parameter of ChatGPT models that governs the randomness and the creativity of the responses. A temperature of 0 essentially that the model will always select the highest probability word conditional on the text, which will eliminate the effect of randomness in the responses and maximize the reproducibility of the results.

And here is ChatGPT 4's response after we replace "_company_name_" with "Humana" and "_headline_" with the text above in the prompt:

> *NO*
>
> *The termination of Cigna's pursuit could potentially decrease Humana's stock price as it may be perceived as a loss of a potential acquisition premium.*

The news headline states that Cigna Group abandoned its pursuit of a merger with Humana and announced a big stock repurchase plan. This news is value-relevant to two health insurance providers: Cigna and Humana. For Humana, the data vendor's proprietary analytics tool gives a positive sentiment score of 0.65, indicating that the news is perceived as favorable for Humana. However, ChatGPT 4 responds that the information is negative for Humana. It reasons that Humana's stock price could drop due to a loss of an acquisition premium caused by the termination of Cigna's bid. The difference in sentiment scores across the models highlights the importance of understanding the context and nuances in prediction tasks.

With the prompting strategy above, we utilize an API provided by OpenAI to prompt ChatGPT and obtain a recommendation for each headline. We use the ChatGPT model version "gpt-4-0314" whose training data stops in September 2021 to ensure that our analysis is an out-of-sample evaluation. We transform it into a numerical "GPT-4 score," where "YES" is mapped to 1, "UNKNOWN" to 0, and "NO" to -1. We present selected descriptive statistics of the sample in Table OA1 of the Online Appendix: (i) the daily stock returns, (ii) the headline length, (iii) the response length, (iv) the GPT-4 score, and the event sentiment score provided by the data vendor. The average GPT-4 score is positive (0.32), with the median being zero, and the event sentiment score shows a similar pattern. Thus, news headlines have a positive tilt. Panel B of the table reports the correlation matrix of these variables.[20]

In addition to analyzing the performance of ChatGPT 4, we examine the capabilities of other more basic models, such as BERT, GPT-1, and GPT-2, and compare their performance

20. In our main prompting strategy, we ask ChatGPT to provide an answer first and then the reasoning. We also tried to ask the model to reason first and then answer in a robustness analysis. Based on a sample of 1,000 randomly selected news headlines, we find that ChatGPT 4's recommendations across the two prompting methods are very similar (see Table OA2 of the Online Appendix).

with that of the more advanced models. We employ a different strategy for the more basic models because those models cannot follow instructions or answer specific questions. For instance, GPT-1 and GPT-2 are auto-complete models. Appendix C of the Online Appendix details the prompts we use for these models.

5 Can ChatGPT Predict Stock Returns?

In this section, we examine LLMs' capabilities in predicting stock price movements. In Section 5.1, we evaluate ChatGPT's performance using a portfolio approach, and in Section 5.2 we employ a regression approach. Section 5.3 compares the performance across different LLMs, and Section 5.4 further analyzes their performance across news types. Section 5.5 studies the potential impact of LLMs on market efficiency.

5.1 Performance of Long-Short Portfolios

We begin by conducting a portfolio analysis to evaluate ChatGPT's ability to predict stock price movements. This involves creating long-short trading strategies based on ChatGPT's scoring of news headlines (buying the stocks with a positive score and selling the ones with a negative score) and analyzing the performance of these portfolios. Due to the differences in the entry time of portfolio formation, we analyze overnight vs. intraday news separately. As over four-fifths of the news headlines are released during nontrading hours (i.e., overnight news), we focus on overnight news in much of our baseline analysis. In particular, if a piece of news is released before 9 a.m. on a trading day, we enter the position at the market opening and exit at the close of the same day. If the news is announced after the market closes, we assume we enter the position at the next opening price and exit at the close of the next trading day. All strategies are rebalanced daily.[21]

Figure 2 plots the cumulative returns over our sample period of four different trading strategies (investing $1) based on ChatGPT 4 without considering transaction costs. These

21. We require at least two firms with positive scores to enter the long leg and similarly with negative scores. Otherwise, we enter just one of the legs. Although we do not use it in the strategy, one could long or short the market portfolio to balance the strategy. Our strategies are not optimized for implementation but rather to show ChatGPT's raw forecasting power.

four strategies include (i) an equal-weighted portfolio that buys companies with good news based on ChatGPT 4 ("Long GPT 4"), (ii) an equal-weighted portfolio that sells companies with bad news based on ChatGPT 4 ("Short GPT 4"), (iii) a self-financing long-short strategy based on ChatGPT 4 ("Long - Short GPT 4"), and (iv) a value-weight market portfolio ("Market Value-Weighted").

[Insert Figure 2 about here]

We find strong evidence of the power of ChatGPT scores in predicting stock returns the next day based on news headlines. For instance, without considering transaction costs, a self-financing strategy that buys stocks with a positive ChatGPT 4 score and sells stocks with a negative ChatGPT 4 score earns an astonishing cumulative return of around 650% from 2021m10 – 2023m12. In contrast, the value-weight market portfolios earn almost zero returns over the same period. The sharp difference in performance demonstrates that ChatGPT can extract valuable information from news headlines and predict stock market reactions after the news announcements. Both the long and short legs contribute to the predictability of ChatGPT. While the long leg delivers a cumulative return of about 70%, the short leg delivers a cumulative return of over 300% during our sample period. Thus, the predictability is stronger among stocks with negative news, consistent with our model's prediction in Proposition 2.2. We note that the large drops (and gains) on specific days in this plot are due to the lack of news on those days combined with an equal-weighting procedure and an adverse prediction, which could be solved with good risk management practices.

Table 1 provides a set of selected statistics of the trading strategies as specified in Figure 2, including the annualized Sharpe ratio, mean daily returns, standard deviation of daily returns, and maximum drawdown. We find that the long-short strategy based on ChatGPT 4 has an average daily return of 38 bps, an annualized Sharpe ratio of 3.28, and the maximum drawdown is -17.42% during our sample period (without considering transaction costs). In all these dimensions, the strategy significantly outperforms the market portfolios or an equal-weight portfolio in all stocks with news the day before (regardless of news direction). When examining the long and short legs separately, we find a clear asymmetry in the predictive power across positive vs negative news, with predictability more pronounced in the short leg as in Figure 2. While the long leg has an average daily return of 9 bps and an annualized

Sharpe ratio of 0.90, the short leg has an average daily return of 29 bps and an annualized Sharpe ratio of 2.12. Again, this evidence is consistent with our model's prediction in Proposition 2.2.

[Insert Table 1 about here]

The analysis in Figure 2 ignores transaction costs, which could be critical given the high turnover ratio of the strategy. In Figure 3, we evaluate the performance of the long-short strategy based on ChatGPT 4 under different transaction cost assumptions: 5, 10, and 20 basis points (bps) per round-trip trade. Assuming a transaction cost of 5 bps round-trip, the strategy still earns a cumulative return of over 300% during our sample period. As we increase the transaction costs to 10 bps round-trip, the cumulative return is still as high as 150%. A transaction cost of 20 bps round-trip makes the strategy unprofitable. Hence, deploying this strategy would need to minimize transaction costs, especially price impact, which is likely infeasible for unsophisticated investors.

[Insert Figure 3 about here]

The long-short strategy in Figure 2 involves all U.S. common stocks with at least one news headline covering the firm. We also analyze the strategy performance by removing small or illiquid stocks from the sample. In particular, as shown in Figure 4, when we remove from the sample stocks with a close price less or equal to $5 and stocks with a market capitalization below the 10th percentile NYSE size breakpoint, the cumulative return of the strategy, without considering transaction costs, is still over 300% during our sample period. This suggests that the predictive power of ChatGPT is not limited to the sample of small stocks but is also present for larger and more liquid stocks.

[Insert Figure 4 about here]

With the analysis above focusing on overnight news, we next form trading strategies based on ChatGPT scores of intraday news. We consider three different entry times when forming portfolios: one-minute post-release, 15 minutes post-release, and 4 p.m. on the news day. Use these entry times, we form three different strategies, each creating an equal-weighted long-short portfolio that buys companies with good news and sells companies with bad news according to ChatGPT 4. The first one enters the position one minute after the news release and exits 15 minutes post-release; the second enters the position 15 minutes

post-release and exits at the market close of the news day; and the third enters the position at the close of the news day and exits at the close of the next trading day.

Figure 5 shows the cumulative returns of the three trading strategies. The first strategy from 1m to 15m post-release earns a cumulative return of around 80% from 2021m10 – 2023m12 without considering transaction costs, the second strategy from 15m post-release to market close earns a cumulative return of over 450%, and the third one from market close to market close of the next day earns a cumulative return of over 350%. Thus, if we enter the position 15 minutes post-release and exit it at the close of the next day, the strategy can earn a cumulative return that is comparable to, if not higher than, that of the overnight news trading strategy shown in Figure 2.[22]

[Insert Figure 5 about here]

5.2 Predictive Regressions Results

In addition to the portfolio analysis, we also use prediction regressions as the second approach to evaluate the performance of different models. Specifically, we the following linear regressions of the next day's stock returns on the ChatGPT score and the sentiment score provided by the data vendor, as follows:

$$r_{i,t+1} = a_i + b_t + \gamma' x_{i,t} + \varepsilon_{i,t+1}, \tag{36}$$

where the dependent variable, $r_{i,t+1}$, is stock i's return over a subsequent trading day after news arrival, $x_{i,t}$ refers to the vector containing the ChatGPT 4 score from assessing stock i's news headlines, and a_i and b_t are firm and date fixed effects, respectively, which account for any observable and unobservable time-invariant firm characteristics and common time-specific factors that could influence stock returns. Standard errors are double clustered by date and firm. Note that this regression analysis is at the news headline level and covers the full sample of headlines (both intraday and overnight news). If the news is released before 9 a.m. on day t or after 4 p.m. on the previous day, $r_{i,t+1}$ is measured from the market

22. There is an additional complication when entering within market hours as the total number of news, and hence the number of positions, is not known in advanced. Although we do not implement it in the paper, and we assume knowledge of the number of news, in practice, a forecasting exercise is needed.

opening of day t to the same day's close. If the news is released after 9 a.m. but before the market close on day t, $r_{i,t+1}$ is measured from the market close price of ay t to the close of the next trading day.

We present the regression results in Table 2. First, we find that the prediction score from ChatGPT 4 has a statistically and economically significant relation with the next-day stock returns. Specifically, the coefficient on ChatGPT 4's score in column (1) is 0.173 with a t-stat of 7.129. A switch from a negative (-1) to a positive (1) prediction score is associated with a 34.6 bps increase in next-day stock return. This evidence corroborates the portfolio analysis in Section 5.1 and further confirms the potential of ChatGPT as a valuable tool for predicting stock market movements based on sentiment analysis. Second, we also compare the performance of ChatGPT with traditional sentiment analysis methods provided by the data vendor (Ravenpack). Our results in columns (2) and (3) show that when controlling for the ChatGPT sentiment scores, the effect of the sentiment score from the data vendor on daily stock returns is attenuated and becomes statistically insignificant. This indicates that the ChatGPT 4 model outperforms existing sentiment analysis methods in forecasting stock market returns.

[Insert Table 2 about here]

In addition, we also examine the predictability of ChatGPT across small vs. non-small stocks by including an indicator variable for stocks below the 10th percentile NYSE market capitalization and its interaction with the ChatGPT 4 score. Our results in column (4) show that the coefficient on the ChatGPT 4 score itself is positive and significant (0.102 with a t-stat of 4.545), suggesting that the predictability of the ChatGPT 4 score is present among both small and large-cap stocks. Thus, the market appears to be underreacting to firm-specific news at the daily frequency we examine, consistent with the evidence documented by the extant literature (e.g., Bernard and Thomas (1989), Chan, Jegadeesh, and Lakonishok (1996), DellaVigna and Pollet (2009), Hirshleifer, Lim, and Teoh (2009), Jiang, Li, and Wang (2021), and Fedyk and Hodson (2023)).

Importantly, we also find that the predictability is more pronounced among the subsample of smaller stocks as shown by the interaction term in column (4). Notably, the coefficient on the ChatGPT 4 score in predicting returns of stocks below the 10th percentile NYSE size

breakpoint is more than six times the magnitude of the one for the remaining sample (i.e., 0.102+0.586 vs 0.102).[23] Therefore, consistent with our theoretical prediction in Proposition 2.2, this evidence highlights the significant role of limits-to-arbitrage in driving the return predictability we document.

Next, we conduct a similar regression analysis for intraday news separately using different return windows. Similar to what we do in Figure 5, we consider three measurement windows for each intraday news headline: (i) return from one minute post-release to 15 minutes post-release, (ii) return from 15 minutes post-release to the market close of the news day, and (iii) return from the market close of the news day to the close of the next trading day. We also consider a combined version of all three returns, corresponding to the return from one minute after the news release to the close of the next trading day. Table 3 reports the results. We find a statistically insignificant coefficient on the ChatGPT 4 score for the return over 1m to 15m post-lease, suggesting that the return potential of the ChatGPT strategy is limited in this short window, consistent with the pattern in Figure 5. In contrast, for the return from 15m post-lease to the news day's close, the coefficient on ChatGPT 4's score is 0.195 with a t-stat of 4.748, while the coefficient is 0.133 with a t-stat of 2.624 for the close-to-close return. Combining the three returns, we get a coefficient of 0.349 with a t-stat of 4.999. It suggests that a switch from a negative (-1) to a positive (1) prediction score for an intraday headline is associated with a 69.8 bps return for the strategy that enters the position one minute post-release and exits it at the close of the next trading day. Thus, it adds significant value to trade intraday news headlines quickly (say within 15 minutes) after the news release to exploit the price movements during the news day.

[Insert Table 3 about here]

In summary, the superiority of ChatGPT in predicting stock market returns can be attributed to its advanced language understanding capabilities, which allow it to capture the nuances and subtleties within news headlines. This enables the model to generate more reliable sentiment scores, leading to better predictions of daily stock market returns. By outperforming traditional sentiment analysis methods, ChatGPT shows potential value in

23. Results are similar if we estimate the regressions separately over the subsamples of small and non-small stocks, rather than using an interaction term specification.

enhancing the performance of quantitative trading strategies and providing a more accurate understanding of market dynamics.

5.3 Comparing Different Large Language Models

In this section, we compare the performance across different LLMs in predicting stock returns. Our theoretical model predicts that only LLMs passing a certain threshold in model size can profitably predict stock returns (see Theorem 1). We empirically test this prediction by contrasting the performance of the basic and more advanced LLMs. Doing so can shed light on whether the return predictability is an emerging capacity of the recent more complex LLMs.

To compare the various language models' performance, we conduct a similar portfolio analysis as in Figure 2 using prediction scores from other LLMs. In particular, we consider the following more basic LLMs (listed based on their release time): (i) GPT-1, (ii) BERT, (iii) BERT Large, (iv) GPT-2, (v) GPT-2 Large, (vi) BART-Large, (vii) DistilBart-MNLI, (viii) GPT-3.5, and (iv) FinBERT. All of these LLMs have been released within the past few years (since 2018), highlighting the rapid advancement and proliferation of AI technologies. The more recent LLMs are generally more complex with a larger number of parameters compared to earlier models. The model sizes, measured by the estimated number of parameters, range from 117 million in the initial GPT-1 model to 175 billion in GPT-3.5, and finally to an astonishing 1.76 trillion in the most advanced model, ChatGPT 4, as detailed in Appendix B of the Online Appendix.

Table 4 reports statistics of long-short portfolios formed based on the assessment scores of news headlines by different models, including the annualized Sharpe ratios, daily average returns, CAPM alpha, Fama-French 5-factor alpha, and average number of stocks in each leg. Our results show a striking pattern—return predictability is an emerging capacity of more complex language models. Scores from advanced but not the most complex models such as BART Large, DistilBart-MNLI, and GPT-3.5 show some predictability but are noticeably weaker compared to the state-of-the-art model, ChatGPT 4. For instance, the average daily returns of the strategies based on DistilBart-MNLI and GPT-3.5 are 17 bps and 34 bps, respectively, compared to 38 bps based on ChatGPT 4. The same pattern remains when we

evaluate their risk-adjusted performance using the CAPM model or the 5-factor model of Fama and French (2015). We find that the strategies all have a low loading on risk factors and their CAPM and 5-factor alphas are almost of the same magnitude as the corresponding daily average returns. Importantly, the annualized Sharpe ratios of the strategies based on DistilBart-MNLI and GPT-3.5 are 1.61 and 1.79, respectively. Below these two LLMs, the next tier includes BART-large and BERT-large, with the annualized Sharpe ratios being 1.24 and 1.12, respectively. All four are significantly less than the Sharpe ratio of 3.28 based on ChatGPT 4. Note that much of the predictability of these LLMs comes from the short leg of the portfolio even though the short leg has fewer stocks on average than the long leg, a pattern also shared by the most complex model ChatGPT 4. When we use the most basic models such as GPT-1, GPT-2, and BERT to assess the news headlines, we do not find their scores correlate significantly positively with subsequent stock returns. The Sharpe ratios of the strategies based on these basic models are all negative. This indicates that most of these models could not predict next-day stock returns based on news headlines with a correct sign.[24]

[Insert Table 4 about here]

In addition, we also conduct pairwise comparisons of the eleven long-short strategies based on different models. For each pair of the daily strategy return series, we examine the intercept (alpha) from a univariate regression using one series to explain the other and then reversing the roles. A significant alpha from such a regression would suggest that the LLM used in the explanatory variable does not perform as well as the model used in the dependent variable. As shown in Table OA4 of the Online Appendix, we find that the strategy based on ChatGPT performs the best and could not be explained by any of the other strategies. Specifically, when we use the return series for ChatGPT 4 as the dependent variable, we obtain a significant daily alpha, ranging from 25 to 39 bps, for the remaining ten models we consider. In general, the ranking based on alphas from pairwise comparison is similar to the ranking based on Sharpe ratios. The next tier is GPT-3.5, DistilBart-MNLI, and Ravenpack, with the remaining models in the bottom tier

24. We find very similar performance ranking across the different models when separately examining small vs. nonsmall stocks, with ChatGPT 4 at the top and models like GPT-1, GPT-2, and BERT at the bottom (see Table OA3 of the Online Appendix).

In a separate analysis, we also compare our ChatGPT-based strategy with the strategy of Jiang, Li, and Wang (2021) that uses the first 15-minute returns to select news to trade. We form long-short portfolios in the same way as in our ChatGPT strategy, except that we use the opening 15-minute returns of overnight news to sort stocks into three groups. Interestingly, we find that the two strategies are independent of each other, with a correlation of less than 0.17 and each series not being able to explain much of the other series (see Table OA5 of the Online Appendix).

In short, the most complex model, ChatGPT 4, exhibits the highest predictability, with the forecasting power of LLMs generally increasing as the model size grows, which aligns well with our theoretical prediction. Therefore, only when AI technologies surpass critical thresholds, can AI performance translate into economic value—profitably predicting stock returns in this case.

5.4 News Complexity and News Types

In this section, we further examine the predictive capabilities of various language models on stock market returns and the complexity of news headlines. Specifically, we categorize headlines based on their Flesch-Kincaid Readability Score, with a daily computed median demarcating them as either low-complexity or high-complexity. This distinction allows us to discern how different models perform when information from news is likely more complicated to understand.

Table 5 reports the results of different models across news complexity. It shows a clear differentiation in model performance. The state-of-the-art model ChatGPT 4 demonstrates the best capability in interpreting complex headlines, with an annualized Sharpe ratio of 1.45, which is higher than that of any other model. Basic models like GPT-1, GPT-2, BERT, and Ravenpack's sentiment analysis show very limited forecasting abilities in high-complexity news. In contrast to the results from high-complexity news, many of these models have some capabilities of interpreting low-complexity news, with Sharpe ratios of strategies based on BART Large, DistilBart-MNLI, GPT-3.5, and Ravenpack's sentiment analysis all above 1.5. This suggests that the sophisticated models have distinct advantages in processing information from complex news sources.

[Insert Table 5 about here]

Furthermore, firm news could come from different sources, with some from news media and others directly released by firms. The last two columns of Table 5 report the annualized Sharpe ratios of different models for different sources of information: news articles vs. press releases. The latter type accounts for over two-thirds of the sample. The most sophisticated AI model ChatGPT 4 produces high Sharpe ratios across both categories (2.55 for news articles and 2.10 for press releases), implying a robust ability to interpret and utilize information from diverse news sources. Its superior performance outshines any other models when analyzing press releases, which could be more susceptible to corporate strategic disclosure given their usual origin from the firms themselves. None of the other models have a Sharpe ratio above 1.2 for press releases. When analyzing news articles, which are third-party communications and may present a less biased view than press releases, many of these models, including BART Large, DistilBart-MNLI, GPT-3.5, and Ravenpack's sentiment analysis, achieve a Sharpe ratio exceeding 1.8. The ability of the most complex model to decipher the subtleties in press releases highlights its resilience in the presence of nuanced signals that the issuing firms could potentially manage or even manipulate (e.g., Cao et al. (2023)).

5.5 Speed of Price Response and Market Efficiency

In a semi-strong efficient market (Fama (1970)), stock prices adjust quickly to new public information. However, as our theoretical model shows, factors such as information capacity constraints and limits to arbitrage could also matter for the speed of news assimilation, and breakthroughs in AI technologies could significantly change information processing and dissemination in the marketplace and, in turn, impact market efficiency. In this section, we first examine the speed of news assimilation over our sample period and then try to examine the potential impact of LLMs on market efficiency.

5.5.1 Speed of Price Response

In this section, we examine the performance of the ChatGPT-based trading strategies in the week after the news arrival to have a more complete picture of how stock prices react

to value-relevant news headlines. By doing so, we aim to shed light on the speed of news assimilation and market efficiency.

We started with forming long-short portfolios based on ChatGPT prediction scores the same way as in Section 5.1. Rather than just hold the long-short strategy for the trading day after the news arrival, we hold it for five full trading days. Figure 6 presents the average one-day holding period returns for the strategy for overnight news and their 95% confidence intervals. Specifically, the plot shows the average returns of the strategy for the news day t=0 (i.e., entering the position at the market opening and exiting at the same day's close) and one-day close-to-close returns for each of the next four days. The results show that ChatGPT 4 scores based on news headlines can predict returns over the next two days, but not afterward. The average daily return for the overnight news strategy is 38 bps on the first trading day after the news arrival and 20 bps on the day after. We repeat the same analysis in Figure 7 for the strategy formed based on intraday news, entering the position 15 minutes post-release and holding it for five full trading days. Again, we find a significant return for the post-news intraday trading session and the day after, but not afterward. Thus, our evidence suggests that information contained in the news is absorbed into market prices in about two days.

[Insert Figures 6 and 7 about here]

To have a more complete picture that covers the period before the news arrives, we also tracked the performance of the long-short portfolios over the few days pre-release. Figure 8 shows the average daily returns before and after the news arrival for the overnight news strategy, separately for the long and short legs. We also show, using the dot in between day t=-1 and t=0 in the plot, the market reactions to the news, i.e., the return from the close of day t-1 to the market opening of day t. The same is done for the intraday news strategy in Figure 9. First, the figures demonstrate that ChatGPT's assessment scores accurately capture the immediate reaction to firm-specific news. Second, the market tends to underreact to the news, and stock prices continue to drift in the same direction as the initial reaction within the next two days but not afterward. This set of evidence aligns well with the underreaction prediction in our theoretical model.

[Insert Figures 8 and 9 about here]

5.5.2 Market Efficiency

GPT models have witnessed phenomenal growth in their adoption since 2022 with the GPT-3.5 model being released in March 2022, and ChatGPT, fine-tuned from a GPT-3.5 model, being launched in November 2022.[25] Consistent with the general trend, Sheng et al. (2024) document a sharp increase in 2022 in the adoption of generative AI by asset management firms such as hedge funds. Given the widespread adoption of the recent LLMs and the forecasting capabilities we document, one important question to examine is that of the potential impact of their adoption on market dynamics such as market efficiency.

Our theoretical model posits that LLMs can increase investors' information processing capabilities and reduce market inefficiencies (see Theorem 2 and Proposition 2.5). Thus, return predictability using ChatGPT's prediction score of news headlines is expected to weaken as LLMs' model size increases substantially and an increasing number of market participants use them for processing information for investment purposes. While testing this hypothesis tightly represents an empirical challenge, we examine whether there are any changes over time in the performance of the long-short strategies formed based on the ChatGPT 4 scores. As shown in Figure 10, there is a clear drop in the performance of the ChatGPT-based strategy over our sample period, over which GPT models' capabilities and adoption skyrocketed. Specifically, the average daily return of the strategy is 65 bps in 2021Q4, 38 bps in 2022, and 30 bps in 2023. The annualized Sharpe ratios show a similar declining pattern in Figure 11: 6.54 in 2021Q4, 3.68 in 2022, and 2.33 in 2023. We view this as suggestive evidence that the recent proliferation of LLM technologies is reducing market underreaction to news and improving market efficiency, which is consistent with our model's prediction if investors are indeed taking advantage of these technologies.

[Insert Figures 10 and 11 about here]

25. See media coverage on the widespread adoption: e.g., "ChatGPT sets record for fastest-growing user base" by Krystal Hu, *Reuters*, February 2, 2023; "ChatGPT Is The Fastest Growing App In The History Of Web Applications" by Cindy Gordon, *Forbes*, February 2, 2023. Statistics from *Similarweb* show that the number of monthly visits to the ChatGPT platform was approximately 152 million in November 2022 and increased all the way to over 1.5 billion by the end of 2023.

6 Interpretability

Traditional machine learning models often prioritize prediction at the expense of interpretability. LLMs, however, offer a unique advantage: they take text as input and provide both a prediction score and an explanation as output. This format could, in theory, allow us to understand the reasoning behind predictions for individual examples by examining the corresponding text behind each prediction. However, the sheer volume of data points makes it impractical to discern global patterns through manual reading. To address this challenge, we propose an interpretability method to understand LLMs' capabilities better.

6.1 Understanding LLMs' Predictions and Performance

Our proposed interpretability method consists of two steps. The first step employs surrogate modeling, a machine-learning technique that utilizes an interpretable model, such as linear regression, to comprehend a more complex one (Molnar (2022)). A surrogate model is a simplified, more interpretable model that approximates the behavior of a complex system—in this case, the LLM. With these simpler models, we can gain insights into LLMs' decision-making process that would be difficult or impossible to discern by directly examining LLMs' internal parameters.[26]

The surrogate approach can be applied to LLMs' direct output, helping us understand what factors influence their predictions. We fit a linear regression model to predict LLMs' scores based on features extracted from the input text (e.g., news headlines) provided to the LLM. This allows us to see which features most strongly impact LLMs' analysis. Alternatively, surrogate modeling can be applied to LLMs' performance metrics to analyze the factors that contribute to their successful or failed predictions. Finally, we can also use surrogate modeling to study the differences in prediction scores or performance across various LLMs, such as GPT-4 and GPT-3.5. This comparative analysis helps identify the factors that drive the improvements or variations across different LLMs.

While surrogate modeling provides a powerful tool for interpreting LLMs, the effective-

26. The quality of the surrogate model can be measured by its approximation power using standard metrics such as the (out-of-sample) R^2 or accuracy.

ness of this approach can be limited by the interpretability of the input features themselves. When working with text data, traditional representation methods like bag-of-words, TF-IDF, or word embeddings often result in high-dimensional, sparse feature spaces that are difficult for people to interpret meaningfully.[27] To address this challenge and further enhance the interpretability of our surrogate models, we introduce a second step: topic modeling, which offers a more intuitive and human-readable text representation. Topics, as coherent clusters of related words, provide a higher-level abstraction that captures the central themes and concepts in the text and makes it easier to understand the underlying patterns and relationships in the data. We adapt the textual factors method of Cong, Liang, and Zhang (2024) and implement it using the state-of-the-art BERTopic technique (Grootendorst (2022)).

The choice of text data for topic modeling, whether the input text of news headlines or LLMs' explanations, provides different insights into LLMs' behavior. When applied to news headlines, the topic model reveals the underlying themes and content structure that influence LLMs' predictions or performance and helps us understand how different news topics or styles affect the model's output. Alternatively, when applied to LLMs' explanations, the topic model uncovers patterns in the model's reasoning process, showing us how it justifies its decisions across various scenarios. We run separate regressions on the topics extracted from news headlines and LLM explanations. Reassuringly, both approaches result in very similar themes.

We primarily employ linear regression for our surrogate modeling due to its interpretability and statistical properties. The discrete nature of topics as features aligns well with the additive nature of linear models, creating an intuitive framework for analysis. This technique allows us to quantify each topic's importance through coefficient magnitudes and determine their reliable impact via statistical significance tests. The topics derived from our topic modeling step serve as semantically meaningful features for the regression model, allowing us to directly link high-level concepts to LLM outputs or performance metrics in a statistically sound manner. While more complex models might capture non-linear relationships, the straightforward nature of linear regression and natural fit with topic-based features make it

27. For instance, a linear regression model fitted on thousands of word features might identify important words but fail to capture higher-level semantic concepts. Similarly, while adequate for many NLP tasks, dense word embeddings lack the intuitive interpretability needed to gain clear insights into LLMs' behavior.

well-suited for our interpretability goals.

Our topic modeling approach yields discrete categories, resulting in topics that function as dummy variables in our regression analysis. This discreteness is particularly advantageous for interpretability. Each topic becomes a binary indicator for the presence or absence of a particular theme. This binary nature facilitates a clear interpretation of the surrogate model's coefficients. For instance, a positive coefficient for a topic indicates that the presence of that theme increases LLMs' score or performance metric, while a negative coefficient suggests the opposite effect. Moreover, the only issue with being too granular is not having enough samples to get statistical significance for topics with a small number of samples.

6.2 Interpretability Results

As explained earlier, LLM prediction scores take the values of -1 (negative), 0 (neutral), or 1 (positive). We construct a measure of prediction performance for each headline as the product of the GPT score and the next day's stock return. Intuitively, if the LLM score is negative and the return is negative, the performance is positive as the LLM's prediction is in the right direction. The performance measure also has the property that the difference between the performance of two models is the difference in their score times the return, similar to going long on one LLM's recommendation and short on the other's.

Tables 6 present the results of several regression models where the dependent variables are the GPT-4 score, GPT-3.5 score, their performance (scores multiplied by returns in percentage), and the differences between these measures. The independent variables are the topics extracted from the news headlines in Panel A and those extracted from LLM explanations in Panel B. Note that only topics significant in predicting GPT-4 performance (G_4*R) or the difference in performance between GPT-4 and GPT-3.5 are displayed for brevity. Reassuringly, results are consistent across the two panels.

[Insert Table 6 about here]

The specific topic modeling algorithm has the advantage that if a headline or explanation does not fit into any of the themes, it leaves it unclassified, and hence, the results of the intercept in the regression can be understood as a baseline effect. There are roughly 40 thou-

sand samples that do not have a clear category.[28] In both specifications, the average baseline GPT-4 score is positive, with an average performance of 18.7 basis points per news headline. GPT-4 performance is higher on average than that of GPT-3.5, and the outperformance is about 10 basis points as shown in the last column of either panel. Hence, the newer model seems better at forecasting stock return movements in general.

On the news headlines, the topic model algorithm is granular enough to distinguish between stock acquisitions from directors, chairpersons, and executives. Executive transactions get, on average, rated too negative by GPT-4, and it results in a decrease in performance of -16 basis points relative to the baseline. Chairman and director transactions are rated close to the baseline, but both result in an outperformance of 30 to 60 basis points. Moreover, director stock transactions are harder to rate for GPT-3.5, and this results in a difference in performance of 33.5 basis points between the two LLMs. On the explanations, both topics related to general insider transactions and chairman transactions are associated with an outperformance of roughly 35 to 70 basis points. Insider transactions are harder for GPT-3.5, and again, there is an outperformance of GPT-4 in comparison by 36.5 basis points.

As expected, share repurchase announcements tend to be rated very positive by both GPT models and more so for GPT-3.5. Both models perform well in their predictions in this category, with an average increase in performance of almost 1 percent. In contrast, ChatGPT is not able to correctly understand the average impact of issuing equity or convertible notes. GPT-4 explains that these funds are issued for firm growth, and they are a positive signal on average, but these categories result in a decrease in performance of 50 basis points. Moreover, on the explanation themes, both models tend to overestimate the effect of awards firms receive on stock prices, resulting in a lower performance.

There are other interesting themes where GPT-4 performs significantly better than its previous version. For example, it does much better at reverse stock splits, which are rated extremely negative by GPT-4 but less so by its predecessor, resulting in an outperformance of around 3.2 percentage points. GPT-4 also does better in very specific industry-related themes, such as fitness and electric vehicles.

28. The details of the topic model algorithm are provided in Appendix E of the Online Appendix. Additionally, the extracted topics can be found in Tables OA6 and OA7 of the Online Appendix.

Interestingly, there are cases where GPT-4 does worse than its previous versions. News headlines and explanations related to cloud strategic partnerships and expansions are rated too positive. GPT-4 tends to think these partnerships have too much of an effect, and it results in a difference in performance of -15 basis points. An inverse pattern occurs with the theme of hotel acquisitions as GPT-4 rates them too low relative to GPT-3.5 and loses in performance comparison. Finally, we control for the similarity between news headlines and explanations as measured by the cosine similarity of their word embeddings. GPT-4 gives explanations very similar to headlines with positive scores, perhaps because positive news appears obvious to GPT-4 from the headlines themselves.

The R^2_{Adj} values indicate that the models explain a reasonable amount of variance in GPT-4 scores (about 35%) and the difference between GPT-4 and GPT-3 scores (around 28%) but very little of the variance in prediction performance measures (less than 1 %). This result suggests that while the topic models somewhat predict LLM scores, they have limited explanatory power for the performance or the accuracy of their predictions. This result is consistent with more basic sentiment analysis methods not working as well as LLMs since, if the performance was predictable to a large degree, we could improve LLMs' predictions using this information. Furthermore, while returns are very noisy, the performance measure should work even better in more stable economic tasks.

In summary, our framework can help interpret and analyze LLMs' behavior on large-scale datasets. The surrogate models provide global views of LLM performance and behavior, while the topic model offers insights into the underlying themes and concepts driving these patterns. This approach also enables tracking and interpreting the progress and differences across different LLMs while maintaining high interpretability.

7 Conclusion

In this study, we investigate the potential of ChatGPT and other large language models in predicting stock market returns using news headlines. We document several empirical findings new to the literature. First, ChatGPT's assessment scores of news headlines can predict subsequent daily stock returns. Its predictability outperforms traditional sentiment analysis

methods. Second, predictability is stronger among smaller stocks and stocks with negative news. Third, more basic LLMs such as GPT-1, GPT-2, and BERT cannot accurately forecast returns. At the same time, strategies based on ChatGPT 4 deliver the highest Sharpe ratio, indicating return predictability is an emerging capacity of complex language models. Fourth, only advanced LLMs can predict returns using more complex news, while more basic models fail at this task. Fifth, we find suggestive evidence that widespread LLM adoption can enhance market efficiency. Importantly, we develop a theoretical model incorporating information capacity constraints, underreaction, limits-to-arbitrage, and LLM technology to explain all of our empirical findings. Finally, we propose a new method to evaluate and understand LLMs' reasoning capabilities. This method can be applied to any LLM-related task to shed light on how these models perform. By demonstrating the value of LLMs in financial economics, our study makes a unique contribution to the growing body of literature on the applications of artificial intelligence and natural language processing in this domain.

Our research has several implications for future studies. First, it highlights the importance of continued exploration and development of LLMs tailored explicitly for the financial industry. As AI-driven finance evolves, more sophisticated models can be designed to improve the accuracy and efficiency of financial decision-making processes. Second, our findings suggest that future research could focus more on understanding the mechanisms through which LLMs derive their predictive power. By identifying the factors contributing to models like ChatGPT's success in predicting stock returns, researchers can develop more targeted strategies for improving these models and maximizing their utility in finance. Third, as LLMs become more prevalent in the financial industry, it is essential to investigate their potential impact on market dynamics, including price formation, information dissemination, and market stability.

Lastly, future studies could explore the integration of LLMs with other machine learning techniques and quantitative models to create hybrid systems that combine the strengths of different approaches. By leveraging the complementary capabilities of various methods, researchers can further enhance the predictive power of AI-driven models in financial economics.

References

Acemoglu, Daron. 2024. "Harms of AI." In *The Oxford Handbook of AI Governance.* Oxford University Press, June.

Acemoglu, Daron, David Autor, Jonathon Hazell, and Pascual Restrepo. 2022. "Artificial Intelligence and Jobs: Evidence from Online Vacancies." *Journal of Labor Economics* 40:293–S340.

Acemoglu, Daron, and Pascual Restrepo. 2022. "Tasks, Automation, and the Rise in U.S. Wage Inequality." *Econometrica* 90, no. 5 (September): 1973–2016.

Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb. 2019. "Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction." *Journal of Economic Perspectives* 33, no. 2 (March): 31–50.

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson. 2024. "Artificial intelligence, firm growth, and product innovation." *Journal of Financial Economics* 151 (January): 103745.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis. 2016. "Measuring economic policy uncertainty." *Quarterly Journal of Economics* 131, no. 4 (November): 1593–1636.

Bernard, Victor L, and Jacob K Thomas. 1989. "Post-earnings-announcement drift: delayed price response or risk premium?" *Journal of Accounting Research* 27:1–36.

Bollerslev, Tim, Sophia Zhengzi Li, and Viktor Todorov. 2016. "Roughing up beta: Continuous versus discontinuous betas and the cross section of expected stock returns." *Journal of Financial Economics* 120 (3): 464–490.

Brown, David P, and Robert H Jennings. 1989. "On technical analysis." *The Review of Financial Studies* 2 (4): 527–551.

Bybee, J Leland. 2023. "The ghost in the machine: Generating beliefs with large language models." *Working Paper.*

Bybee, Leland, Bryan T. Kelly, Asaf Manela, and Dacheng Xiu. 2023. "Business News and Business Cycles." *SSRN Electronic Journal* (September).

Calomiris, Charles W., and Harry Mamaysky. 2019. "How news and its context drive risk and returns around the world." *Journal of Financial Economics* 133, no. 2 (August): 299–336.

Cao, Sean, Wei Jiang, Baozhong Yang, and Alan L. Zhang. 2023. "How to Talk When a Machine Is Listening: Corporate Disclosure in the Age of AI." *The Review of Financial Studies* 36, no. 9 (August): 3603–3642.

Chan, Louis KC, Narasimhan Jegadeesh, and Josef Lakonishok. 1996. "Momentum strategies." *The Journal of Finance* 51 (5): 1681–1713.

Chen, Yifei, Bryan T Kelly, and Dacheng Xiu. 2023. "Expected returns and large language models." *Available at SSRN 4416687.*

Cohen, Lauren, Christopher Malloy, and Quoc Nguyen. 2020. "Lazy Prices." *Journal of Finance* 75 (3): 1371–1415.

Cong, Lin William, Tengyuan Liang, and Xiao Zhang. 2024. "Textual Factors: A Scalable, Interpretable, and Data-driven Approach to Analyzing Unstructured Information." *SSRN Electronic Journal.*

Dávila, Eduardo, and Cecilia Parlatore. 2018. "Identifying price informativeness." *NBER Working Paper 25210.*

———. 2021. "Trading costs and informational efficiency." *The Journal of Finance* 76 (3): 1471–1539.

De Long, J Bradford, Andrei Shleifer, Lawrence H Summers, and Robert J Waldmann. 1990. "Noise trader risk in financial markets." *Journal of Political Economy* 98 (4): 703–738.

DellaVigna, Stefano, and Joshua M Pollet. 2009. "Investor inattention and Friday earnings announcements." *The Journal of Finance* 64 (2): 709–749.

Eisfeldt, Andrea L, Gregor Schubert, and Miao Ben Zhang. 2023. "Generative AI and firm values." NBER Working Paper 31222.

Fama, Eugene F. 1970. "Efficient capital markets: A review of theory and empirical work." *The Journal of Finance* 25:383–417.

Fama, Eugene F., and Kenneth R. French. 2015. "A five-factor asset pricing model." *Journal of Financial Economics* 116, no. 1 (April): 1–22.

Fedyk, Anastassia, and James Hodson. 2023. "When can the market identify old news?" *Journal of Financial Economics* 149, no. 1 (July): 92–113.

Freyberger, Joachim, Andreas Neuhierl, and Michael Weber. 2020. "Dissecting Characteristics Nonparametrically." *The Review of Financial Studies* 33 (5): 2326–2377.

Garcia, Diego. 2013. "Sentiment during Recessions." *The Journal of Finance* 68, no. 3 (June): 1267–1300.

Gromb, Denis, and Dimitri Vayanos. 2010. "Limits of arbitrage." *Annuual Review of Financial Economics* 2 (1): 251–275.

Grootendorst, Maarten. 2022. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure" (March).

Grossman, Sanford J, and Joseph E Stiglitz. 1980. "On the impossibility of informationally efficient markets." *American Economic Review* 70 (3): 393–408.

Grundy, Bruce D, and Maureen McNichols. 1989. "Trade and the revelation of information through prices and direct disclosure." *The Review of Financial Studies* 2 (4): 495–526.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu. 2020. "Empirical Asset Pricing via Machine Learning." *The Review of Financial Studies* 33 (5): 2223–2273.

Hansen, Anne Lundgaard, and Sophia Kazinnik. 2023. "Can ChatGPT Decipher Fedspeak?" *SSRN Electronic Journal* (March).

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2018. "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach." *The Quarterly Journal of Economics* 133, no. 2 (May): 801–870.

Hirshleifer, David, Sonya Seongyeon Lim, and Siew Hong Teoh. 2009. "Driven to distraction: Extraneous events and underreaction to earnings news." *The Journal of Finance* 64 (5): 2289–2325.

Hoberg, Gerard, and Gordon Phillips. 2016. "Text-Based Network Industries and Endogenous Product Differentiation." *Journal of Political Economy* 124 (5): 1423–1465.

Jegadeesh, Narasimhan, and Di Wu. 2013. "Word power: A new approach for content analysis." *Journal of Financial Economics* 110 (3): 712–729.

Jiang, Fuwei, Joshua Lee, Xiumin Martin, and Guofu Zhou. 2019. "Manager sentiment and stock returns." *Journal of Financial Economics* 132, no. 1 (April): 126–149.

Jiang, Hao, Sophia Zhengzi Li, and Hao Wang. 2021. "Pervasive underreaction: Evidence from high-frequency data." *Journal of Financial Economics* 141, no. 2 (August): 573–599.

Jiang, Wei, Yuehua Tang, Rachel Jiqiu Xiao, and Vincent Yao. 2023. "Surviving the FinTech disruption." NBER Working Paper 28668.

Ke, Shikun, José Luis Montiel Olea, and James Nesbit. 2019. "A Robust Machine Learning Algorithm for Text Analysis." *Working Paper.*

Ke, Zheng, Bryan T Kelly, and Dacheng Xiu. 2019. "Predicting Returns with Text Data." *University of Chicago, Becker Friedman Institute for Economics Working Paper.*

Ko, Hyungjin, and Jaewook Lee. 2023. "Can Chatgpt Improve Investment Decision? From a Portfolio Management Perspective." *SSRN Electronic Journal.*

Kogan, Leonid, Dimitris Papanikolaou, Lawrence DW Schmidt, and Bryan Seegmiller. 2023. "Technology and labor displacement: Evidence from linking patents with worker-level data." NBER Working Paper 31846.

Korinek, Anton. 2023. "Generative AI for economic research: Use cases and implications for economists." *Journal of Economic Literature* 61 (4): 1281–1317.

Kyle, Albert S. 1985. "Continuous auctions and insider trading." *Econometrica,* 1315–1335.

———. 1989. "Informed Speculation with Imperfect Competition." *The Review of Economic Studies* 56, no. 3 (July): 317–355.

Lerner, Josh, Amit Seru, Nicholas Short, and Yuan Sun. 2024. "Financial Innovation in the Twenty-First Century: Evidence from US Patents." *Journal of Political Economy* 132 (5): 1391–1449.

Lo, Andrew W. 2002. "The statistics of Sharpe ratios." *Financial Analysts Journal* 58 (4): 36–52.

Lopez-Lira, Alejandro. 2023. "Risk Factors That Matter: Textual Analysis of Risk Disclosures for the Cross-Section of Returns." *SSRN Electronic Journal.*

Manela, Asaf, and Alan Moreira. 2017. "News implied volatility and disaster concerns." *Journal of Financial Economics* 123, no. 1 (January): 137–162.

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

McInnes, Leland, John Healy, and Steve Astels. 2017. "hdbscan: Hierarchical density based clustering." *Journal of Open Source Software* 2, no. 11 (March): 205.

McInnes, Leland, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. "UMAP: Uniform Manifold Approximation and Projection." *Journal of Open Source Software* 3, no. 29 (September): 861.

Molnar, Christoph. 2022. *Interpretable Machine Learning.* 2nd ed.

Noy, Shakked, and Whitney Zhang. 2023. "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence." *SSRN Electronic Journal* (March).

Rapach, David E, Jack K Strauss, and Guofu Zhou. 2013. "International stock return predictability: What is the role of the united states?" *Journal of Finance* 68 (4): 1633–1662.

Sheng, Jinfei, Zheng Sun, Baozhong Yang, and Alan Zhang. 2024. "Generative AI and Asset Management." *Available at SSRN 4786575.*

Shleifer, Andrei, and Robert W Vishny. 1997. "The limits of arbitrage." *The Journal of Finance* 52 (1): 35–55.

Tetlock, Paul C. 2007. "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." *The Journal of Finance* 62, no. 3 (June): 1139–1168.

———. 2011. "All the News That's Fit to Reprint: Do Investors React to Stale Information?" *The Review of Financial Studies* 24, no. 5 (May): 1481–1512.

Tetlock, Paul C., Maytal Saar-Tsechansky, and Sofus Macskassy. 2008. "More Than Words: Quantifying Language to Measure Firms' Fundamentals." *Journal of Finance* 63, no. 3 (June): 1437–1467.

Van Nieuwerburgh, Stijn, and Laura Veldkamp. 2010. "Information Acquisition and Under-Diversification." *The Review of Economic Studies* 77, no. 2 (April): 779–805.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you need." *Advances in Neural Information Processing Systems* 2017-Decem:5999–6009.

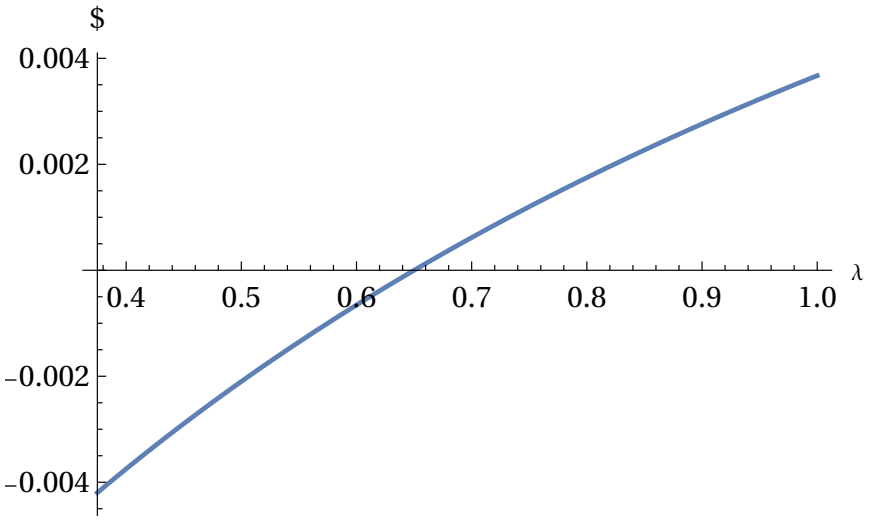
Verrecchia, Robert E. 1982. "Information Acquisition in a Noisy Rational Expectations Economy." *Econometrica* 50, no. 6 (November): 1415.

Webb, Michael. 2019. "The Impact of Artificial Intelligence on the Labor Market." *SSRN Electronic Journal* (November).

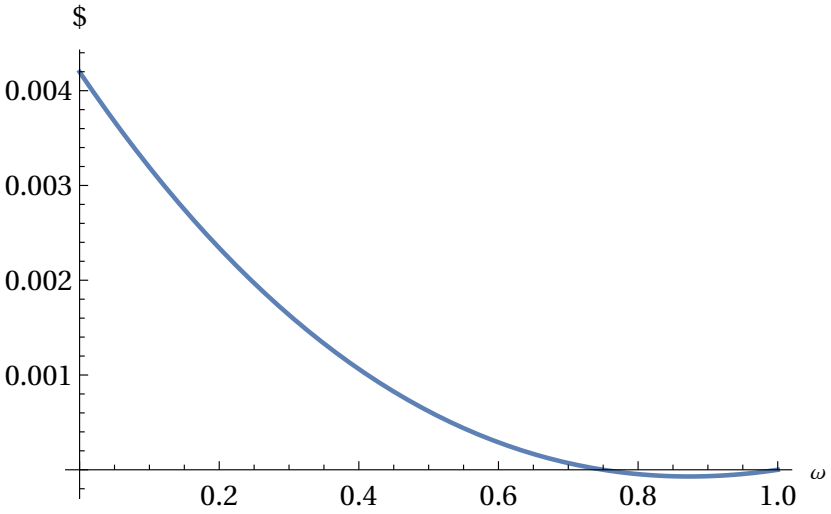
Wu, Shijie, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. "BloombergGPT: A Large Language Model for Finance" (March).

Xie, Qianqian, Weiguang Han, Yanzhao Lai, Min Peng, and Jimin Huang. 2023. "The Wall Street Neophyte: A Zero-Shot Analysis of ChatGPT Over MultiModal Stock Movement Prediction Challenges" (April).

Figure 1: Expected Profitability of LLM Strategy



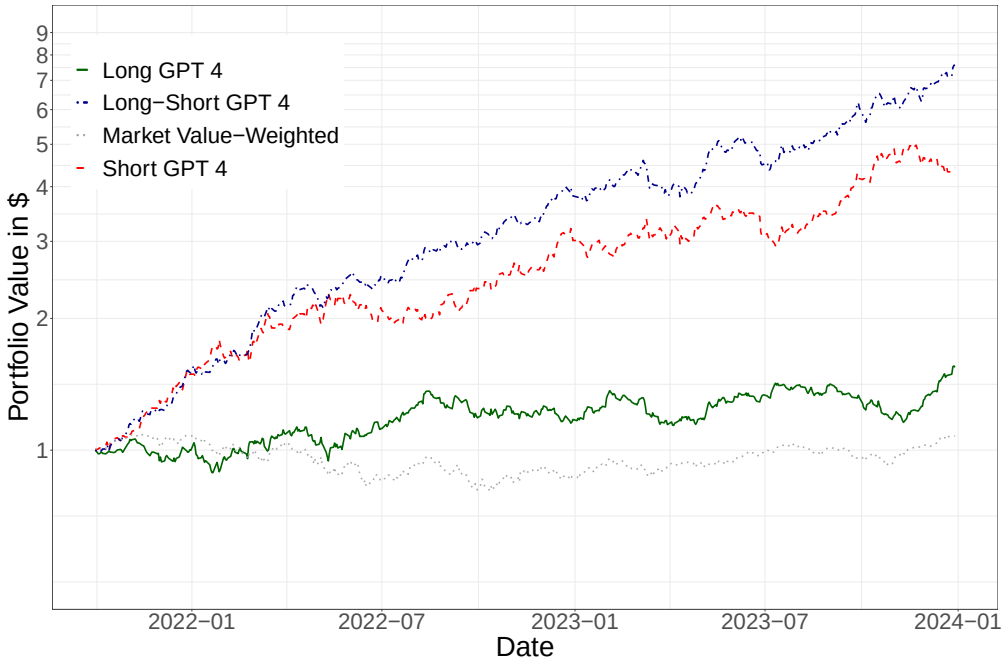
(a)



(b)

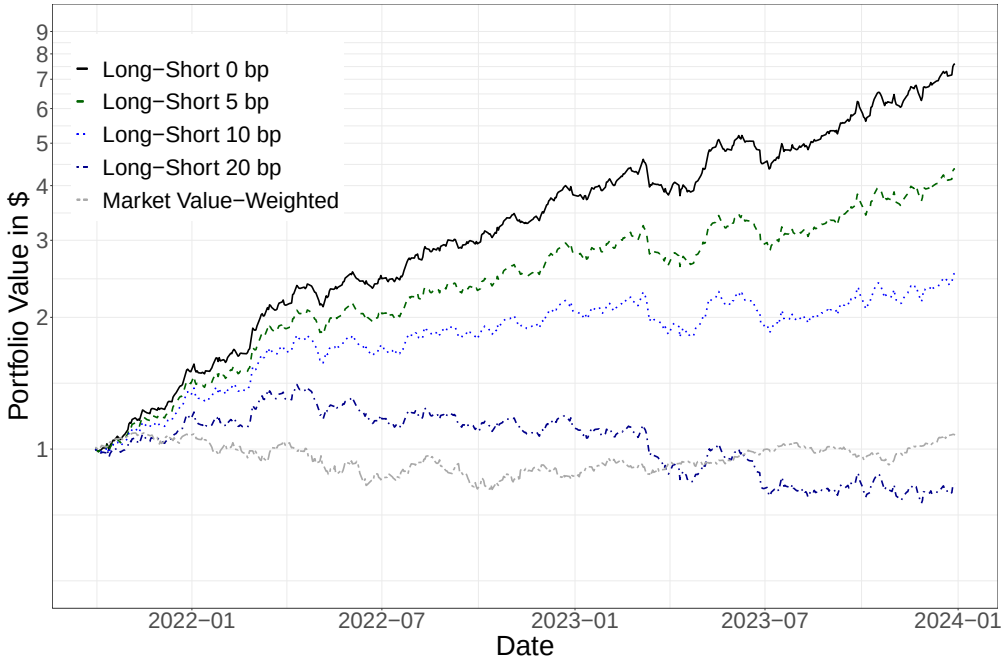
This figure presents how the expected dollar profitability of LLM-based strategies (without considering trading costs) differ when changing the following parameter at a time: the information capacity of the LLM technology (λ) in the first panel and the information capacity of the inattentive agents (ω) in the second panel. The values used in the figure are the proportion of attentive agents ($\pi_A = 0.5$), the proportion of inattentive agents ($\pi_I = 0.5$), the precision of the dividend ($\tau_D = 1$), the precision of the signal ($\tau_S = 1$), the signal realization ($s = 2$), the mean dividend ($\bar{d} = 1$), the attentive agents' information capacity ($\gamma_A = 0.75$), the inattentive agents' relative information capacity ($\omega = 0.5$, only in the first panel), the precision of nonfundamental demand ($\tau_U = 1$), the risk aversion ($\alpha = 1$), and the LLM information capacity ($\lambda = 0.7$, only in the second panel).

Figure 2: Cumulative Returns of Investing $1 (Without Transaction Costs)



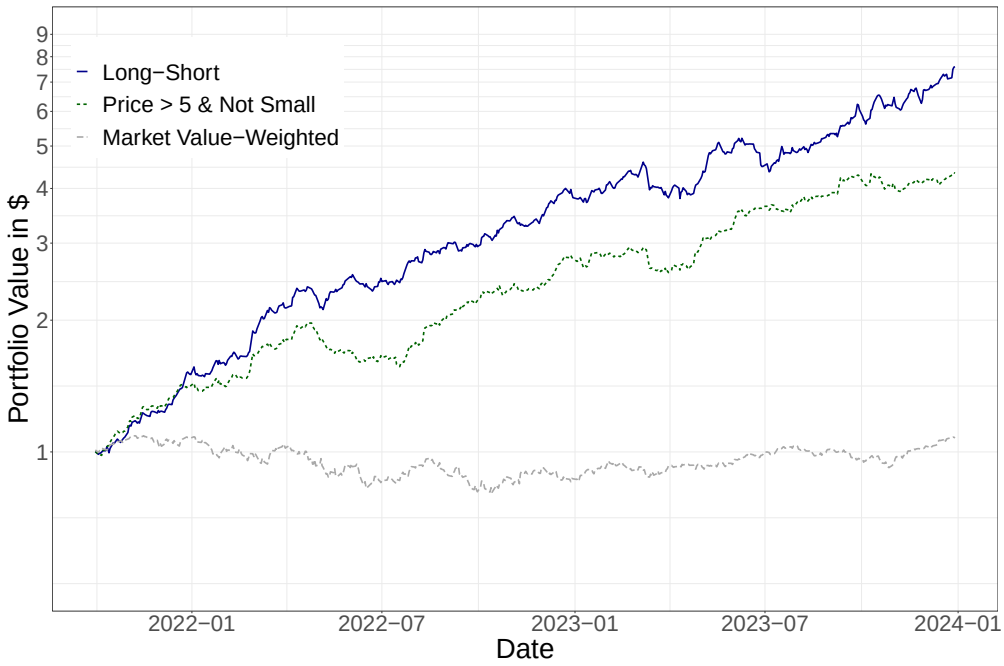
This figure shows the performance of trading strategies based on ChatGPT 4's prediction scores of overnight news, without considering transaction costs. If a piece of news is released before 9 a.m. on a trading day, we enter the position at the market opening and exit at the close of the same day. If the news is announced after the market closes, we assume we enter the position at the next opening price and exit at the close of the next trading day. All the strategies are rebalanced daily. The green line corresponds to an equal-weighted portfolio that buys companies with good news, according to ChatGPT 4. The red line corresponds to an equal-weighted portfolio that short-sells companies with bad news, according to ChatGPT 4. The blue line corresponds to an equal-weighted zero-cost portfolio that buys companies with good news and short-sells companies with bad news, according to ChatGPT 4. The grey line corresponds to a value-weighted market portfolio.

Figure 3: Cumulative Returns of Investing $1 With Different Transaction Costs



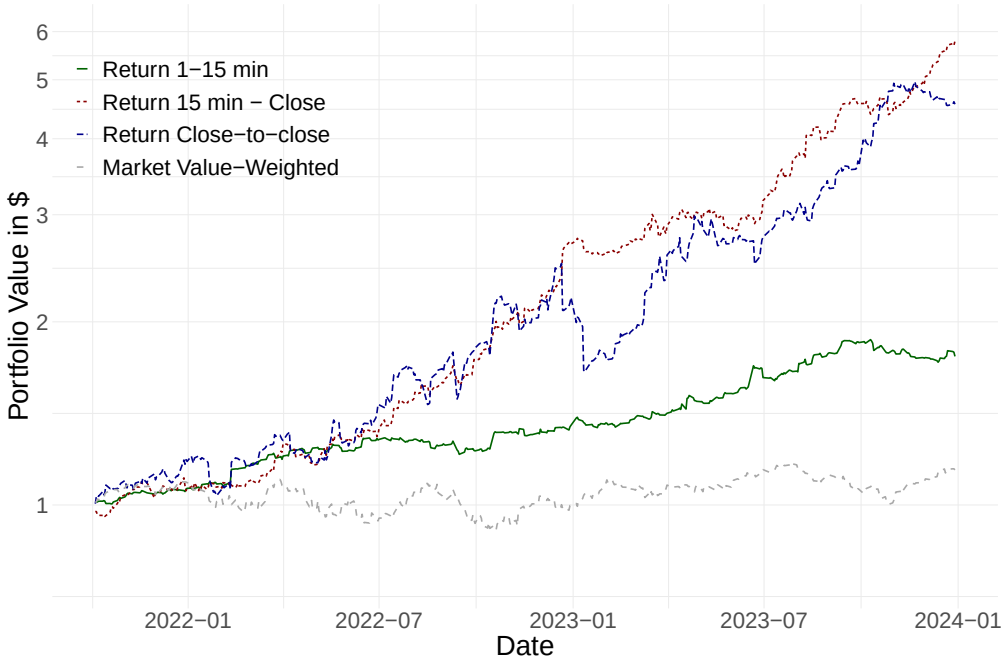
This figure shows the performance of trading strategies based on ChatGPT 4's prediction scores of overnight news, considering different transaction costs. If a piece of news is released before 9 a.m. on a trading day, we enter the position at the market opening and exit at the close of the same day. If the news is announced after the market closes, we assume we enter the position at the next opening price and exit at the close of the next trading day. All the strategies are rebalanced daily. All the strategies are rebalanced daily. The black line corresponds to an equal-weighted zero-cost portfolio that buys companies with good news and short-sells companies with bad news, according to ChatGPT 4, with zero transaction costs. The dark green line corresponds to the same equal-weighted zero-cost portfolio with a 5 bps round trip cost. The light blue line corresponds to the same equal-weighted zero-cost portfolio with a 10 bps round trip cost. The dark blue line corresponds to the same equal-weighted zero-cost portfolio with a 20 bps round trip cost. The grey line corresponds to a value-weighted market portfolio without transaction costs.

Figure 4: Cumulative Returns of Investing $1 With Different Sample Restrictions



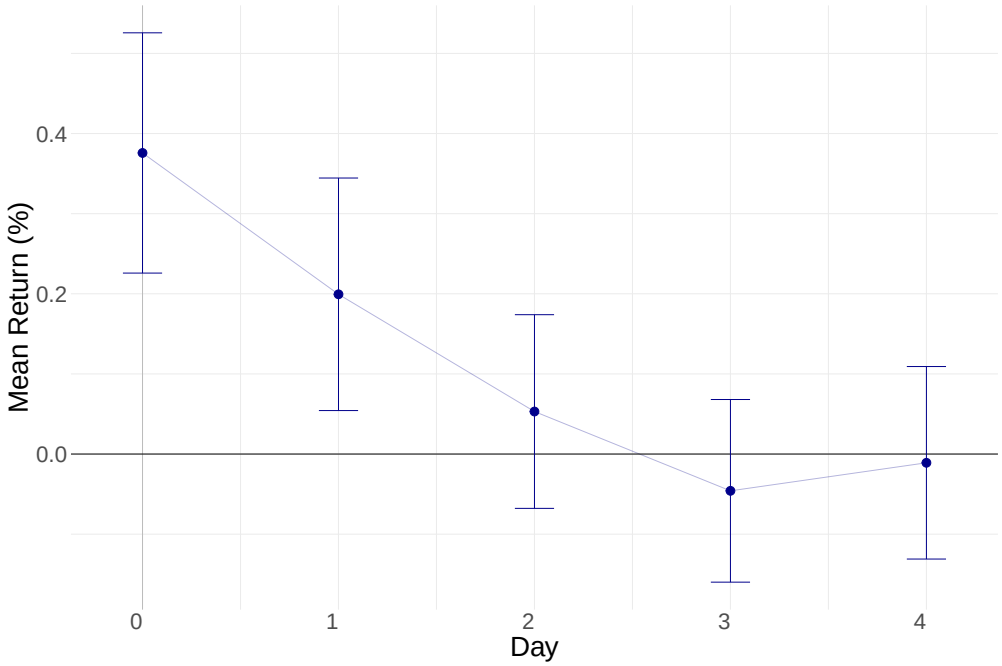
This figure shows the performance of trading strategies based on ChatGPT 4's prediction scores of overnight news over different samples, without considering transaction costs. If a piece of news is released before 9 a.m. on a trading day, we enter the position at the market opening and exit at the close of the same day. If the news is announced after the market closes, we assume we enter the position at the next opening price and exit at the close of the next trading day. All the strategies are rebalanced daily. All the strategies are rebalanced daily. The blue line corresponds to an equal-weighted zero-cost portfolio that buys companies with good news and short-sells companies with bad news according to ChatGPT 4 for the full sample of U.S. common stocks. The green line corresponds to the same equal-weighted zero-cost portfolio but over the sample of stocks with a close price greater than $5 on the prior day *and* market capitalization above the 10th percentile NYSE market cap breakpoint. The grey line corresponds to a value-weighted market portfolio without transaction costs.

Figure 5: Cumulative Returns of Investing $1 in the Long-Short Strategy for Intraday News



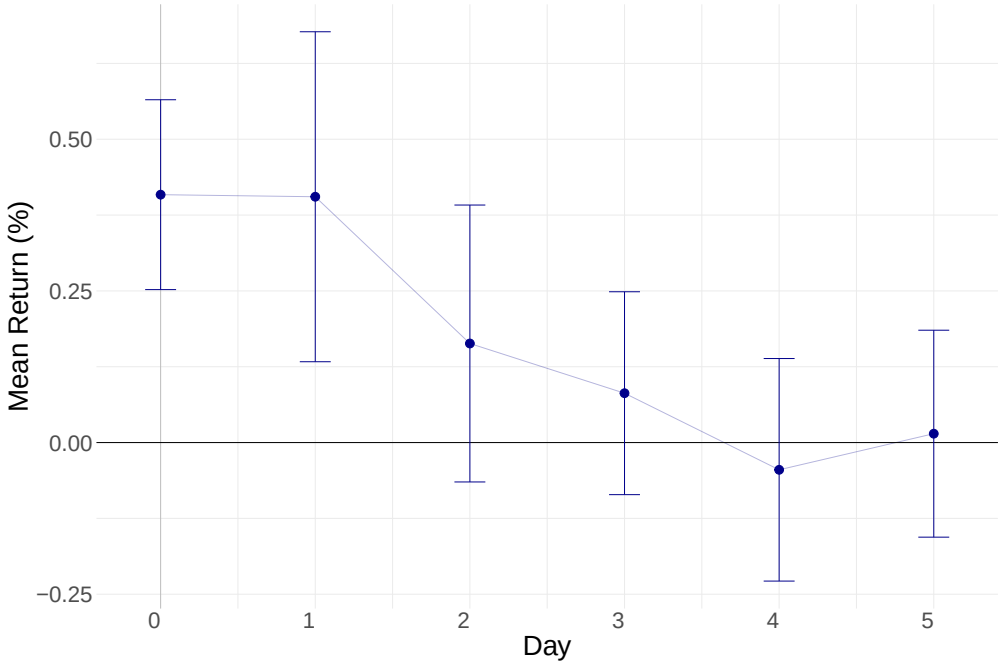
This figure shows the performance of trading strategies based on ChatGPT 4's prediction scores of intraday news, without considering transaction costs. For intraday news released between 9:30 a.m. and 4 p.m. of trading day t, we form three strategies with different entry times, each creating an equal-weighted long-short portfolio that buys companies with good news and short-sells companies with bad news according to ChatGPT 4. The first strategy (the green line) enters the position one minute after the news announcement and exits 15 minutes after the news announcement on day t; the second one (the red line) enters the position 15 minutes after the news announcement and exits at the market close of day t; the third one (the blue line) enters the position at the market close of day t and exits at the market close of day $t+1$. All the strategies are rebalanced daily. The grey line corresponds to a value-weighted market portfolio without transaction costs.

Figure 6: Returns of Overnight News Strategy Over Event Time



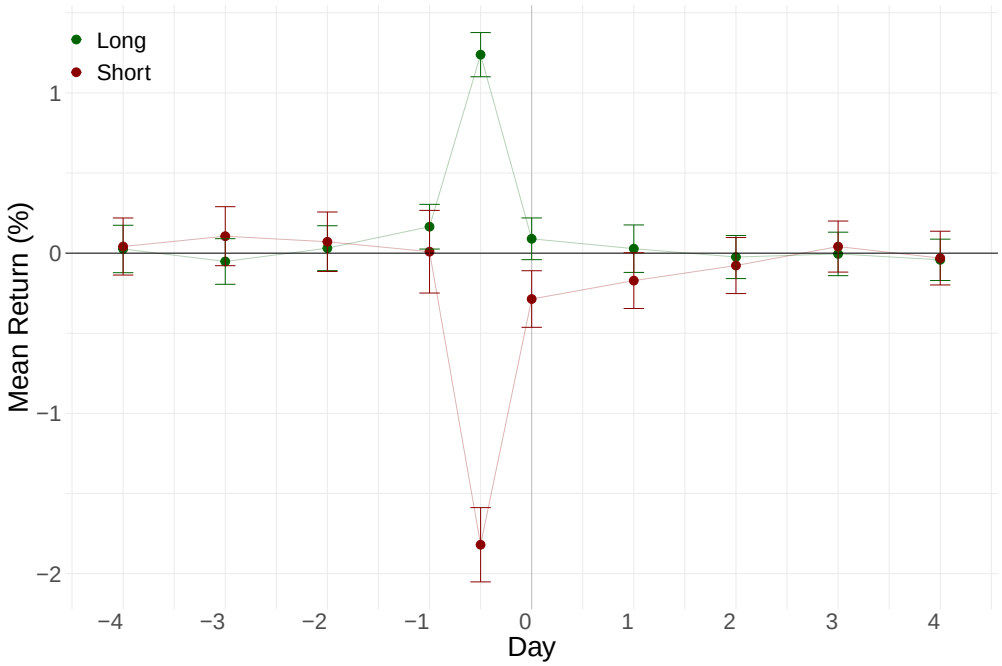
This figure presents the average one-day holding period returns for the strategy using overnight news and their 95% confidence intervals. For overnight news released before 9 a.m. on trading day t but after 4 p.m. of day t-1, we enter the position at the market opening of day t, creating a long-short portfolio that buys companies with good news and short-sells companies with bad news according to ChatGPT 4, and hold it for five trading days. The plot shows the average returns of the strategy for day t=0 (i.e., entering the position at the market opening of day t and exiting at the same day's close) and one-day close-to-close returns for each of the next four days, day t+1 through t+4.

Figure 7: Returns of Intraday News Strategy Over Event Time



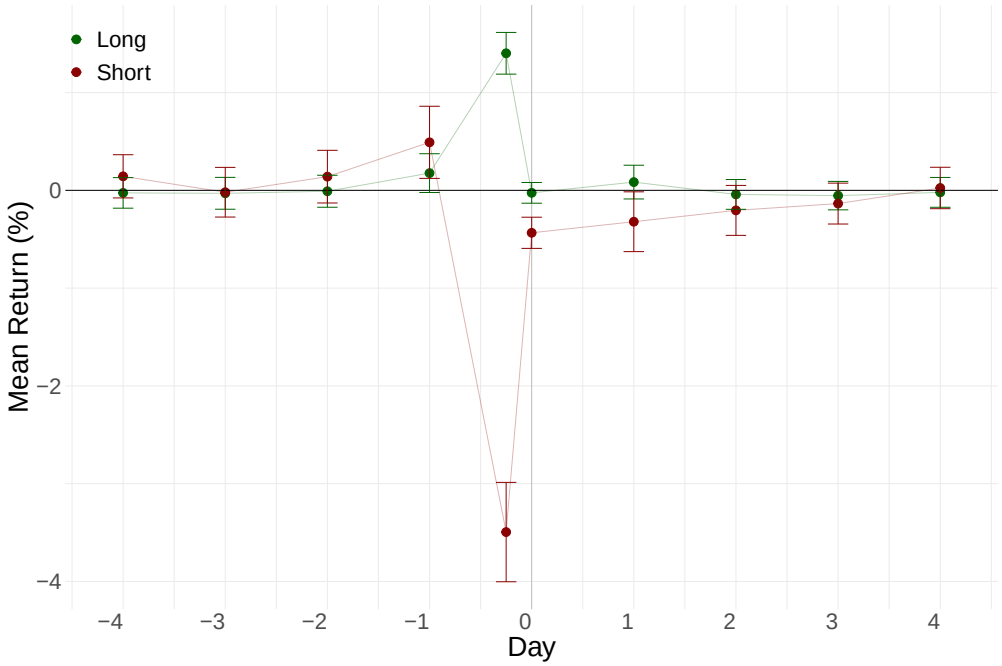
This figure presents the average one-day holding period returns for the strategy using intraday news and their 95% confidence intervals. For intraday news released between 9:30 a.m. and 4 p.m. of trading day t, we enter the position 15 minutes after the news announcement on day t, creating a long-short portfolio that buys companies with good news and short-sells companies with bad news according to ChatGPT 4, and hold it for five full trading sessions. The plot shows the average returns of the strategy for day t=0 (i.e., entering the position 15 minutes after the news announcement on day t and exiting at the same day's close) and one-day close-to-close returns for each of the next five days, day t+1 through t+5.

Figure 8: Overnight News Returns: Before and After the Release Time



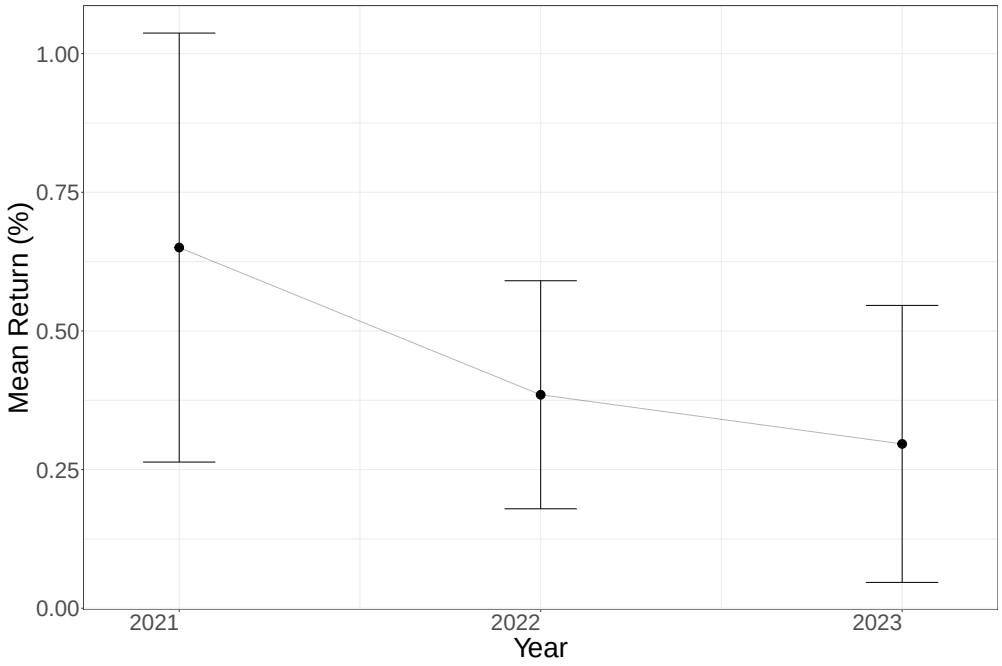
This figure presents the average one-day holding period returns over event time for the strategy using overnight news and their 95% confidence intervals. For overnight news released before 9 a.m. on trading day t but after 4 p.m. of day t-1, we enter the position at the market opening of day t based on ChatGPT 4 scores and hold it for five trading days. The plot shows the average returns of each leg of the strategy for day t=0 (i.e., entering the position at the market opening of day t and exiting at the same day's close) and one-day close-to-close returns for the few days before and after day t — day t-4 through t-1 and t+1 through t+4. The dot in between day t=-1 and t=0 in the plot corresponds to market reactions to the news, i.e., the return from the close of day t-1 to the market opening of day t. The green line corresponds to an equal-weighted portfolio that buys companies with good news according to ChatGPT 4. The red line corresponds to an equal-weighted portfolio that short-sells companies with bad news according to ChatGPT 4.

Figure 9: Intraday News Returns: Before and After the Release Time



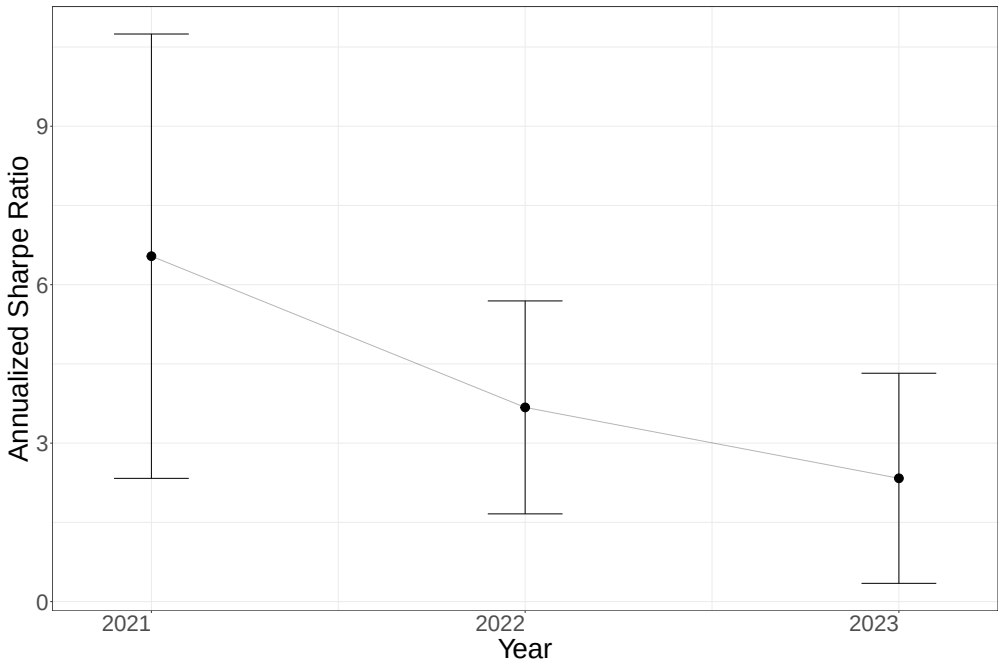
This figure presents the average one-day holding period returns over event time for the strategy using intraday news and their 95% confidence intervals. For intraday news released between 9:30 a.m. and 4 p.m. of trading day t, we enter the position 15 minutes after the news announcement on day t, creating a long-short portfolio that buys companies with good news and short-sells companies with bad news according to ChatGPT 4, and hold it for five full trading sessions. The plot shows the average returns of each leg of the strategy for day t=0 (i.e., entering the position 15 minutes after the news announcement on day t and exiting at the same day's close) and one-day close-to-close returns for the few days before and after day t — day t-4 through t-1 and t+1 through t+4. The dot in between day t=-1 and t=0 in the plot corresponds to market reactions to the news, i.e., the return from the close of day t-1 to 15 minutes after the news announcement on day t. The green line corresponds to an equal-weighted portfolio that buys companies with good news according to ChatGPT 4. The red line corresponds to an equal-weighted portfolio that sells companies with bad news according to ChatGPT 4.

Figure 10: Year-by-Year Performance of Overnight News Strategy: Average Daily Returns



This figure shows the year-by-year performance of the same long-short strategy based on ChatGPT 4 as in Figure 2. If a piece of news is released before 9 a.m. on a trading day, we enter the position at the market opening and exit at the close of the same day. If the news is announced after the market closes, we assume we enter the position at the next opening price and exit at the close of the next trading day. All the strategies are rebalanced daily. This figure presents the average daily returns of the strategy for three subperiods of our sample (2021Q4, 2022, and 2023) and their 95% confidence intervals.

Figure 11: Year-by-Year Performance of Overnight News Strategy: Sharpe Ratios



This figure shows the year-by-year Sharpe ratios of the same long-short strategy based on ChatGPT 4 as in Figure 2. If a piece of news is released before 9 a.m. on a trading day, we enter the position at the market opening and exit at the close of the same day. If the news is announced after the market closes, we assume we enter the position at the next opening price and exit at the close of the next trading day. All the strategies are rebalanced daily. This figure presents the annualized Sharpe ratios of the strategy for three subperiods of our sample (2021Q4, 2022, and 2023) and their 95% confidence intervals calculated following Lo (2002).

Table 1: Descriptive Statistics of Various Portfolios

This table reports several statistics of the different trading strategies as specified in Figure 2, including the annualized Sharpe ratio, mean daily returns, standard deviation of daily returns, and maximum drawdown. The strategies include (i) the long, short, and long-short strategy based on ChatGPT 4, (ii) equal-weight and value-weight market portfolios, and (iii) an equal-weight portfolio in all stocks with news the day before (regardless of news direction).

	LS GPT 4	Long	Short	Market EW	Market VW	All News EW
Ann. Sharpe Ratio	3.28	0.90	2.12	-0.28	0.27	-0.57
Daily Mean (%)	0.38	0.09	0.29	-0.02	0.02	-0.05
Daily Std. Dev. (%)	1.82	1.58	2.14	1.11	1.19	1.50
Max Drawdown (%)	-17.42	-18.62	-19.13	-31.01	-25.73	-46.22

Table 2: Regression of Next Day Returns on Prediction Scores

This table reports the results from regressions of the form $r_{i,t+1} = a_i + b_t + \gamma' x_{i,t} + \varepsilon_{i,t+1}$, where $r_{i,t+1}$ is the next day's return in percentage points. The analysis covers all U.S. common stocks with at least one news headline (intraday or overnight) covering the firm. If a piece of news is released before 9 a.m. on a trading day or after 4 p.m. of the previous day, we enter the position at the market opening and exit at the close of the same day. If the news is released after 9 a.m. but before the market close, we enter the position at the market close price of the same day and exit at the close of the next trading day. The terms a_i and b_t are firm and time fixed effects, respectively. $x_{i,t}$ corresponds to the vector containing the prediction score from ChatGPT 4 and, for comparison purposes, the event sentiment score from Ravenpack. *Small* is an indicator variable for stocks below the 10th percentile NYSE market capitalization the previous day. The corresponding t-statistics are in parentheses. Standard errors are double clustered by date and firm. All models include firm and time fixed effects. The table reports results for all U.S. common stocks with at least one news headline covering the firm.

	(1)	(2)	(3)	(4)
GPT-4 Score	0.173***	0.186***		0.102***
	(7.129)	(6.086)		(4.545)
RavenPack		−0.055	0.130**	
		(−0.873)	(2.590)	
GPT-4 Score*Small				0.586***
				(4.982)
Small				0.378+
				(1.664)
Num.Obs.	134 129	134 129	134 129	134 129
R2 Adj.	0.094	0.094	0.094	0.095
Std.Errors	by: date & permno	by: date & permno	by: date & permno	by: date & permno
FE: date	X	X	X	X
FE: permno	X	X	X	X

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Table 3: Regression of Intraday News Returns on Prediction Scores

This table reports the results of running regressions of the form $r_{i,t+1} = a_i + b_t + \gamma' x_{i,t} + \varepsilon_{i,t+1}$, where $r_{i,t+1}$ is the strategy return in percentage points based on intraday news, and $x_{i,t}$ refers to the prediction score from ChatGPT 4. For intraday news released between 9:30 a.m. and 4 p.m. of trading day t, we form different strategies with different entry times, each creating a long-short portfolio that buys companies with good news and sells companies with bad news according to ChatGPT 4. The first strategy ($Return1 - 15min$) enters the position one minute after the news announcement and exits 15 minutes after the news; the second one ($Return15min - Close$) enters the position 15 minutes after the news and exits at the close of day t; the third one ($ReturnClose - to - close$) enters the position at the close of day t and exits at the next day's close. $TotalReturn$ cumulates the returns of the three strategies, i.e., entering at one minute after the news and exiting at the close of the next day. The corresponding t-statistics are in parentheses. Standard errors are double clustered by date and firm. All models include firm and time-fixed effects.

	Total Return	Return 1-15 min	Return 15 min - Close	Return Close-to-close
GPT-4 Score	0.349***	0.035	0.195***	0.133**
	(4.999)	(1.628)	(4.748)	(2.624)
Num.Obs.	22 272	22 272	22 272	22 272
R2 Adj.	0.268	0.214	0.243	0.261
Std.Errors	by: date & permno	by: date & permno	by: date & permno	by: date & permno
FE: date	X	X	X	X
FE: permno	X	X	X	X

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Table 4: Average Next Day's Return by Prediction Score

This table reports a list of statistics for the portfolios formed based on different models using overnight news headlines (over 80% of the full sample). The column $Sharpe_{LS}$ presents the annualized Sharpe ratio of the long-short portfolio for each model, and the next four columns contain the daily average returns in percentage points (0.10 corresponds to 0.10%) for the long-short, long, neutral, and short portfolios, respectively. We also report the average number of stocks in the long (N_+) and in the short (N_-) legs. Column α_M shows the daily alpha with respect to the CAPM model, t α_M is the t-statistic, and R^2_M is the R-sq. from the CAPM model in percentage points (0.10 corresponds to 0.10%). Columns α_{FF5}, t α_{FF5}, and R^2_{FF5} show the same but with respect to the 5-factor model of Fama and French (2015). We provide an overview of the different LLMs in Appendix B of the Online Appendix. The table reports results for all U.S. common stocks with at least one news headline covering the firm.

Model	Sharpe$_{LS}$	μ_{LS}	μ_+	μ_0	μ_-	N_+	N_-	α_M	t α_M	R^2_M	α_{FF5}	t α_{FF5}	R^2_{FF5}
GPT-4	3.28	0.38	0.09	-0.22	-0.29	70	20	0.38	4.84	0.09	0.37	4.71	0.54
DistilBart-MNLI-12-1	1.61	0.17	-0.03	-0.02	-0.21	115	16	0.17	2.42	0.57	0.18	2.44	1.35
GPT-3.5	1.49	0.26	0.05	-0.09	-0.21	49	6	0.27	2.25	0.41	0.26	2.16	1.37
Ravenpack	1.39	0.19	-0.00	-0.06	-0.20	53	16	0.19	2.16	0.01	0.19	2.18	0.52
BART-Large	1.24	0.14	-0.03	-0.04	-0.17	112	19	0.14	1.87	0.49	0.15	2.06	1.63
BERT-Large	1.12	0.18	-0.06	-0.06	-0.24	122	2	0.18	1.69	2.58	0.20	1.91	4.31
GPT-1	-0.31	-0.03	-0.05	-0.14	-0.01	101	18	-0.03	-0.46	0.03	-0.03	-0.45	0.29
GPT-2	-0.31	-0.04	-0.05	-0.08	-0.01	82	19	-0.04	-0.48	0.01	-0.04	-0.46	0.43
FinBERT	-0.43	-0.09	-0.15	-0.05	-0.06	22	8	-0.09	-0.65	0.01	-0.09	-0.65	1.27
BERT	-0.61	-0.07	-0.08	-0.05	-0.00	34	0	-0.08	-1.16	21.28	-0.05	-0.71	34.11
GPT-2-Large	-0.93	-0.17	-0.09	-0.05	0.08	53	11	-0.17	-1.41	0.20	-0.18	-1.47	0.68

Table 5: Sharpe Ratios by News Complexity and News Type

This table reports the annualized Sharpe ratio of the long-short portfolio implied by different models. The first column reports results using all news. The second column reports results using only low-complexity news, and the third column reports results using only high-complexity news. A headline is non-complex if it is below the median score of the Flesch-Kincaid Readability Score (Flesch and Kincaid 1975) and complex otherwise. The median is computed each day and separately for news during regular market hours and outside of regular market hours. In the last two columns, we analyze the two types of news separately: news articles and press releases. The fourth column reports results using only news articles, and the fifth column reports results using only press releases. We provide an overview of the different LLMs in Appendix B of the Online Appendix.

Model	All	Low Complexity	High Complexity	News Articles	Press Releases
GPT-4	3.28	2.60	1.45	2.55	2.10
GPT-3.5	1.79	2.61	0.21	1.92	0.99
DistilBart-MNLI-12-1	1.61	1.53	0.22	1.81	0.49
Ravenpack	1.39	2.17	0.52	2.94	0.82
BART-Large	1.24	1.81	0.45	1.87	1.12
BERT-Large	1.12	-0.29	1.43	0.51	0.75
GPT-1	-0.31	-1.32	0.01	-0.13	0.26
GPT-2	-0.31	-0.45	-0.23	1.17	-0.44
FinBERT	-0.43	-0.66	0.28	-0.30	0.25
BERT	-0.61	-0.17	-0.49	0.54	-0.38
GPT-2-Large	-0.93	-0.30	-1.03	0.08	-0.80

Table 6: Interpretability

This table reports the results of our interpretability analysis, showing the impact of various topics extracted from news headlines or LLM explanations on GPT-4 and GPT-3.5 scores and their performance in predicting stock returns. The first column lists the news topics identified by our first-step topic modeling algorithm and the number of headlines (in thousands) for each topic. Subsequent columns show coefficients from linear regressions of predicting GPT-4 score (G_4), GPT-4 performance (G_4*R), GPT-3.5 score (G_3), GPT-3.5 performance (G_3*R), the difference between GPT-4 and GPT-3.5 scores (ΔG), and the difference in their performance (ΔG*R). Coefficients in the score columns indicate how each topic influences the LLM's predictions, with positive values suggesting a more positive assessment of the news headlines. Coefficients in performance columns show how well these assessment scores predict subsequent returns, with positive values indicating better prediction accuracy. The intercept represents the baseline effect for unclassified news headlines and explanations, while the "Similarity News Explanations" row shows the impact of similarity between news headlines and model explanations. This analysis focuses only on non-neutral GPT scores (87,699 data points). Panel A presents the results for topics extracted from news headlines, while Panel B shows those extracted from LLM explanations. Standard errors are double clustered by firm and date but are omitted for brevity.

Panel A. News Headlines

	G_4	G_4*R	G_3	G_3*R	ΔG	ΔG*R
Intercept 39.72K	0.298***	0.187**	0.205***	0.088+	0.093***	0.099**
Executive Stock Transactions 1.59K	−0.224***	−0.165*	−0.197***	−0.062	−0.027	−0.104
Chairman Stock Transactions 1.05K	0.038	0.361*	0.027	0.359**	0.011	0.002
Strategic Cloud Partnerships 1.03K	0.396***	−0.111	0.385***	0.043	0.010	−0.154***
Director Stock Transactions 0.75K	−0.058	0.674***	−0.027	0.339**	−0.031	0.335**
Share Repurchase Announcements 0.53K	0.498***	1.097***	0.600***	1.193***	−0.103***	−0.096
Convertible Senior Notes Offerings 0.46K	0.443***	−0.530*	−0.164***	−0.129	0.608***	−0.401
Hotel Acquisition and Sales 0.22K	0.146+	0.045	0.150	0.279	−0.004	−0.234**
Reverse Stock Splits Announced 0.2K	−1.137***	4.545***	−0.447***	1.310*	−0.690***	3.235**
EV Market Dynamics 0.16K	−0.551***	0.283	−0.387***	−0.018	−0.165***	0.301**
Fitness Equipment 0.12K	−0.411*	−0.409*	−0.352+	−0.947***	−0.058***	0.537*
Similarity News Explanations	0.535***	−0.164	0.368***	−0.042	0.166***	−0.123+
R^2(%)	34.6	0.3	15.4	0.1	27.9	0.3

Panel B: LLM Explanations

	G_4	G_4*R	G_3	G_3*R	ΔG	ΔG*R
Intercept 40.46K	0.299***	0.186**	0.208***	0.089+	0.091***	0.098**
Chairman Share Transactions 1.03K	0.041	0.345*	0.025	0.369**	0.016	−0.024
Cloud Partnerships Boost Revenue 1.02K	0.392***	−0.075	0.383***	0.076	0.009	−0.151***
Insider Confidence in Company 0.74K	−0.061	0.710***	−0.030	0.344**	−0.031	0.365**
Stock Repurchase Confidence 0.55K	0.484***	0.995***	0.589***	1.101***	−0.105***	−0.106
Capital Raising for Growth 0.49K	0.437***	−0.498*	−0.144***	−0.143	0.581***	−0.355
Award Impact on Stocks 0.32K	0.472***	−0.404*	0.319***	−0.374*	0.154***	−0.030
Strong Performance Boosts Stocks 0.24K	0.161*	0.081	0.126	0.282	0.035	−0.200*
Impact of Reverse Stock Split 0.2K	−1.140***	4.552***	−0.452***	1.318*	−0.688***	3.234**
Stock Price Fluctuations 0.16K	−0.555***	0.290	−0.392***	−0.009	−0.163***	0.299**
Similarity News Explanations	0.539***	−0.176	0.371***	−0.057	0.168***	−0.118+
N	87,699	87,699	87,699	87,699	87,699	87,699
R^2(%)	34.5	0.3	15.2	0.1	27.8	0.3

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

ONLINE APPENDIX

Can ChatGPT Forecast Stock Price Movements?
Return Predictability and Large Language Models

This Version: September 2024

Appendix A: Additional Tables

Table OA1: Descriptive Statistics

Panel A of this table reports selected descriptive statistics of the daily stock returns in percentage points, the headline length, the response length, the GPT-4 score (1 if ChatGPT says YES, 0 if UNKNOWN, and -1 if NO), and the event sentiment score provided by the data vendor. Panel B reports the correlation between daily stock returns in percentage points, the headline length, the response length, the GPT score, and the event sentiment score.

Panel A. Summary Statistics

	Mean	SD	min	P25	Median	P75	Max	N
Daily Return (%)	−0.05	5.84		−1.98	0	1.83	358.69	134,129
Headline Length	74.55	27.90	19	55	68	88	701	134,129
GPT-4 Response Length	135.26	37.68	0	106	129	159	301	134,129
GPT-4 Score	0.32	0.74	−1	0	0	1	1	134,129
Ravenpack Sentiment Score	0.13	0.36	−1	0	0	0.49	1	134,129

Panel B. Correlations

	Daily Return (%)	Headline Length	GPT-4 Resp. Length	GPT-4 Score	Ravenpack Sent. Score
Daily Return (%)	1	.	.	.	.
Headline Length	−0.006	1	.	.	.
GPT-4 Resp. Length	0.005	0.247	1	.	.
GPT-4 Score	0.020	0.153	0.247	1	.
Ravenpack Sent. Score	0.006	0.074	0.107	0.494	1

Table OA2: ChatGPT Predictions: Answering First or Reasoning First

This table tabulates the distribution (%) of predictive recommendations by different ways of prompting ChatGPT 4 for 1,000 news headlines randomly selected from our sample. In the rows, we use the standard prompt mentioned in Section 4 of the paper and ask the model, ChatGPT 4, to provide the answer first and then provide the reasoning for the recommendation. In the columns, we ask the model to reason first and then provide a recommendation.

	Reason First		
Answer First	NO	UNKNOWN	YES
NO	16.80	0.20	0.00
UNKNOWN	1.00	34.70	0.90
YES	0.60	3.80	42.00

Table OA3: Average Next Day's Return by Prediction Score: Small vs NonSmall Stocks

This table repeats the analysis of different LLMs in Table 4 for small and non-small stocks separately. Panel A analyzes the sample of small stocks (below the 10th percentile NYSE market capitalization), and Panel B analyzes the remaining non-small stocks. All the statistics are the same as in Table 4. We provide an overview of the different LLMs in Appendix B of the Online Appendix.

Panel A. Small Stocks

Model	Sharpe_{LS}	μ_{LS}	μ_+	μ_0	μ_-	N_+	N_-	α_M	t α_M	R^2_M	α_{FF5}	t α_{FF5}	R^2_{FF5}
GPT-4	1.01	0.32	-0.21	-0.64	-0.53	12	2	0.32	1.42	0.10	0.34	1.48	1.24
Ravenpack	0.34	0.13	-0.45	-0.50	-0.58	9	2	0.12	0.49	0.95	0.17	0.65	2.39
GPT-3.5	-0.27	-0.11	-0.31	-0.47	-0.20	9	1	-0.11	-0.44	0.90	-0.07	-0.26	3.18
GPT-2	-0.39	-0.24	-0.51	-0.29	-0.27	13	3	-0.24	-0.59	0.04	-0.22	-0.57	0.36
BART-Large	-0.45	-0.12	-0.44	-0.13	-0.32	20	2	-0.12	-0.68	0.05	-0.08	-0.46	2.43
DistilBart-MNLI-12-1	-0.62	-0.17	-0.44	-0.03	-0.27	20	2	-0.17	-0.96	0.03	-0.13	-0.75	2.07
GPT-1	-0.89	-0.33	-0.41	-0.54	-0.09	17	3	-0.33	-1.33	0.14	-0.31	-1.26	0.69
GPT-2-Large	-0.90	-0.59	-0.49	-0.38	0.10	8	2	-0.60	-1.36	0.11	-0.59	-1.35	0.81
FinBERT	-1.02	-0.41	-0.66	-0.31	-0.25	3	0	-0.42	-1.56	0.77	-0.38	-1.41	2.13
BERT-Large	-1.60	-0.32	-0.44	-0.18	-0.13	21	0	-0.33	-2.59	8.00	-0.29	-2.28	13.42
BERT	-1.64	-0.57	-0.57	-0.33	0.00	5	0	-0.58	-2.49	1.21	-0.53	-2.34	7.24

Panel B. Non-Small Stocks

Model	Sharpe_{LS}	μ_{LS}	μ_+	μ_0	μ_-	N_+	N_-	α_M	t α_M	R^2_M	α_{FF5}	t α_{FF5}	R^2_{FF5}
GPT-4	3.36	0.32	0.15	0.03	-0.18	57	17	0.32	5.03	0.06	0.32	4.93	0.34
BART-Large	2.97	0.27	0.12	-0.01	-0.15	87	16	0.27	4.30	0.00	0.27	4.44	0.94
DistilBart-MNLI-12-1	2.81	0.26	0.11	0.08	-0.16	90	14	0.26	4.05	0.00	0.27	4.13	1.05
Ravenpack	1.85	0.21	0.11	0.08	-0.09	41	14	0.21	2.78	0.31	0.20	2.68	0.64
GPT-3.5	1.64	0.25	0.16	0.04	-0.09	38	5	0.25	2.49	0.31	0.23	2.39	1.54
BERT-Large	1.53	0.23	0.06	0.10	-0.16	102	2	0.22	2.35	3.80	0.25	2.64	6.18
BERT	0.68	0.07	0.07	0.07	-0.00	28	0	0.07	1.05	28.04	0.10	1.68	38.99
GPT-1	0.32	0.02	0.08	0.04	0.05	82	15	0.02	0.47	0.22	0.02	0.33	1.17
GPT-2	0.04	0.00	0.07	0.05	0.07	66	15	0.00	0.07	0.01	0.01	0.20	0.82
FinBERT	-0.27	-0.05	0.04	0.06	0.08	17	7	-0.05	-0.40	0.00	-0.05	-0.42	1.21
GPT-2-Large	-0.80	-0.10	0.05	0.06	0.15	44	9	-0.10	-1.22	1.30	-0.11	-1.32	2.70

Table OA4: Alphas from Comparing Strategies based on Different LLMs

In this table, we run univariate regressions to conduct pairwise comparisons of the long-short strategies based on different models in Table 4. For each pair of the daily strategy return series, we run univariate regressions using one series to explain the other and then reversing the roles. Each cell represents an univariate regression, reporting the alpha (intercept) and its statistical significance. The first row lists the dependent variable for each regression and the first column lists the regressor for each regression. We provide an overview of the different LLMs in Appendix B of the Online Appendix. Note: + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001.

Variable	GPT-4	GPT-3.5	GPT-1	GPT-2	GPT-2-Large	Ravenpack	BART-Large	DistilBart-MNLI-12-1	BERT	BERT-Large	FinBERT
GPT-4		-0.09	-0.05	-0.04	-0.09	0.03	-0.09	-0.07	-0.08	0.02	-0.35**
GPT-3.5	0.27***		-0.04	-0.02	-0.14	0.14	0.07	0.10	-0.07	0.13	-0.23
GPT-1	0.38***	0.27*		-0.04	-0.17	0.19*	0.14	0.18*	-0.07	0.18	-0.09
GPT-2	0.38***	0.26*	-0.03		-0.16	0.2*	0.14	0.17*	-0.08	0.18	-0.1
GPT-2-Large	0.36***	0.25*	-0.04	-0.03		0.21*	0.13	0.17*	-0.07	0.18	-0.07
Ravenpack	0.32***	0.20	-0.03	-0.05	-0.21		0.09	0.12	-0.08	0.15	-0.15
BART-Large	0.29***	0.17	-0.03	-0.03	-0.14	0.14		0.07	-0.08	0.13	-0.17
DistilBart-MNLI-12-1	0.25***	0.13	-0.04	-0.03	-0.15	0.11	0.00		-0.08	0.09	-0.20
BERT	0.38***	0.27*	-0.03	-0.05	-0.16	0.20*	0.15	0.17*		0.2	-0.08
BERT-Large	0.34***	0.22	-0.04	-0.03	-0.16	0.17	0.11	0.13	-0.10		-0.13
FinBERT	0.39***	0.3**	-0.03	-0.04	-0.16	0.20*	0.16*	0.19**	-0.07	0.19	

Table OA5: Comparison of Strategies based GPT-4 and Jiang, Li, and Wang (2021)

This table compares the returns of our GPT-4-based strategy with that of the strategy of Jiang, Li, and Wang (2021) that uses the first 15-minute returns to select overnight news to trade. We use GPT-4 to classify overnight news into three groups (positive, neutral, and negative) each trading day and form equal-weighted zero-cost portfolios long in stocks with news rated positive and short in those with negative news. To facilitate comparison, we use the opening 15-minute returns of the same set of overnight news to sort stocks into three groups and create long-short portfolios of the two extreme groups. Panel A reports descriptive statistics of the daily portfolio returns in percentage points. Panel B reports the correlation between the daily returns of the two strategies. Panel C reports the univariate regression results using one of the two daily return series to explain the other.

Panel A. Summary Statistics

Variable	Mean	Std Dev	Min	P25	Median	P75	Max
Strategy - GPT-4 (%)	0.376	1.818	-10.576	-0.560	0.355	1.217	11.461
Strategy - Open 15m Ret (%)	0.283	1.252	-4.709	-0.394	0.170	0.972	6.977

Panel B. Correlations

	Strategy - GPT 4	Strategy - Open 15m Ret
Strategy - GPT-4	1	
Strategy - Open 15m Ret	0.167***	1
	(<.001)	

Panel C. Alphas

Variable	DV = Strategy - Open 15m Ret	DV = Strategy - GPT-4
Intercept	0.234***	0.301***
	(4.40)	(3.94)
Strategy - GPT-4	0.129***	
	(4.43)	
Strategy - Open 15m Ret		0.263***
		(4.43)
Obs.	558	558
Adj. R-sq	0.032	0.032

Table OA6: News Topics

This table shows the topics generated using news headlines. The column "OpenAI" contains the label generated by ChatGPT. The column "KeyBERT" contains the representative words. The column "Representative Document" contains the document closest to the center of the cluster. Topic -1 corresponds to all the headlines not classified into clusters.

Topic	Count	OpenAI	KeyBERT	Representative Document
-1	39721	Baseline	NA	NA
0	10548	Therapeutics Offerings and Clinical Trials	offering common stock, common stock	Viking Therapeutics Announces Pricing of $250 Million Public Offering of Common Stock
1	6958	Executive Stock Surrenders	surrender, national bank holding	VP Underwood Surrenders 71 Of Dril-Quip Inc
2	4784	Executive and Board Appointments	pharmaceutical appoints, board director	M.D.C. HOLDINGS APPOINTS JANICE SINDEN TO ITS BOARD OF DIRECTORS
3	3514	Cash Dividend Declarations	cash dividend, declares cash dividend	Rent-A-Center, Inc. Declares Quarterly Cash Dividend of $0.34 for the Third Quarter of 2022
4	1594	Executive Stock Transactions	acquires company, acquires	VP Huntley Acquires 518 Of AT&T Inc
5	1361	Record Revenue and Guidance Announcements	report result announces, announces record result	PGTI Reports Record Third Quarter 2023 Results
6	1287	Profitable Sales and Earnings	post higher revenue, profit higher	Amphenol Posts Higher 4Q Profit, Sales on Strong Customer Demand
7	1047	Chairman Stock Transactions	company chmn, industry chmn	Chmn O'Sullivan Acquires 16207 Of CTS Corp
8	1032	Strategic Cloud Partnerships	google cloud, partner google	Crayon: Crayon expands partnership with Google Cloud
9	960	CFO Stock Transactions	cfo, holding cfo	CFO Rehard Acquires 5506 Of Regal Rexnord Corp
10	836	Credit Rating Affirmations	fitch, affirms	Fitch Affirms Bio-Rad at 'BBB'; Outlook Stable
11	752	Renewable Energy and EV Charging Agreements	blink charging, blink	Brunswick Corporation signs Virtual Power Purchase Agreement with Vesper Energy to offset electricity demand with renewable solar energy
12	751	Director Stock Transactions	company dir, industry dir	Dir Masterson Buys 3357 Of Bogota Financial Corp
13	722	Share Price Movements & Earnings	share rise, share drop earnings	Tilly's Shares Fall 10% on Gloomy 1Q Outlook
14	698	Upcoming Investor Conferences	participate investor conference, upcoming investor conference	IAS to Participate at Upcoming Investor Conferences
15	675	Credit Rating Updates	pgr revise, pgr	S&PGR Raises Nasdaq Rtgs To 'BBB+' From 'BBB'; Outlook Stable
16	554	Earnings vs. Consensus Estimates	eps cent factset, factset consensus cent	Kimberly-Clark Q3 adj. EPS $1.74, up 24%; FactSet consensus $1.60
17	533	Share Repurchase Announcements	share repurchase program, announces share repurchase	Gamer Pakistan Announces Share Repurchase Program
18	524	CEO Stock Transactions	buy ceo, biopharma ceo	CEO Derrickson Buys 5000 Of BSQUARE Corp
19	512	Shareholder Merger Investigations	announces acquisition, completes merger	SHAREHOLDER ALERT: Rigrodsky Law, P.A. Announces Investigation of 26 Capital Acquisition Corp. Merger

Topic	Count	OpenAI	KeyBERT	Representative Document
20	492	Stock Price Target Adjustments	price target raised, price target	J.Jill started at buy with $35 stock price target at B. Riley
21	490	Executive Stock Gifts	gift, company dir	Dir Belling Gifts 150 Of RH
22	459	Convertible Senior Notes Offerings	senior note offering, convertible note	8x8, Inc. Prices $137.5 Million of Convertible Senior Notes
23	397	Airlines' Growth and Mergers	united airline, american airline	JetBlue To End JV With American Airlines To 'focus' On Spirit Merger
24	351	Financial Services Presentations	present goldman sachs, participate goldman sachs	Discover Financial Services to Present at the Goldman Sachs 2022 US Financial Services Conference
25	327	Institutional Stock Transactions	holder, holding buy	Holder Feinglass Acquires 22182 Of P10 Inc
26	324	Award-Winning Achievements	product award, innovation award	8x8 Wins Stevie(R) Awards in 2023 American Business Awards(R) for Customer Service
27	316	Global Industrial Conferences	baird global industrial, global industrial conference	Carrier to Present at Baird's 2023 Global Industrial Conference
28	284	Stock Reactions to Earnings	stock rally, stock rise	Bill.com stock drops 20% as weak revenue forecast overshadows earnings beat
29	271	Earnings Release Calls	earnings release conference, earnings conference	Financial Institutions, Inc. Schedules Third Quarter 2021 Earnings Release and Conference Call
30	243	Retail Store Grand Openings	announces opening, grand opening	Floor & Decor Announces Grand Opening of Springfield, New Jersey Store
31	223	Hotel Acquisition and Sales	hilton grand vacation, hilton grand	VP Corbin Jr Acquires 4047 Of Hilton Grand Vacations Inc
32	214	Shareholder Class Action Alerts	lawsuit filed, shareholder alert	AXSOME THERAPEUTICS, INC. (NASDAQ: AXSM) SHAREHOLDER CLASS ACTION ALERT: Bernstein Liebhard LLP Announces that a Securities Class Action Lawsuit Has Been Filed Against Axsome Therapeutics, Inc. (NASDAQ: AXSM)
33	209	Legal Announcements on Corporate Investigations	announces investigation, investigation	Lifshitz Law PLLC Announces Investigation of BREZ, MCAE, NSEC, and IDFB
34	201	Reverse Stock Splits Announced	reverse stock split, stock split	22nd Century Announces 1-for-15 Reverse Stock Split
35	191	Sidoti Virtual Investor Presentations	virtual investor conference, investor conference	Amesite Inc. to Present at Sidoti Virtual Investor Conference December 8-9, 2021
36	190	Record Earnings and Loan Growth	earnings, record earnings	Citizens Community Bancorp, Inc. Reports Earnings Of $0.41 Per Share in 2Q22; Net Interest Margin Expands to 3.46%; Originated Loans Up 6.0% From Prior Quarter
37	182	Credit Card Partnerships	credit card, mastercard	Wells Fargo Partners with Bilt Rewards and Mastercard to Issue the First Credit Card that Earns Points on Rent payments without the Transaction Fee
38	174	Declares Cash Dividend	declares cash dividend, cash dividend declares	SEE Declares Quarterly Cash Dividend
39	172	Defense and Aerospace Contracts	awarded contract, receives contract	Sarcos Defense Awarded $1M Contract by U.S. Army
40	168	Major Tech Antitrust Mergers	activision blizzard, activision	EU to Probe Microsoft's $75 Billion Deal for Activision Blizzard More Deeply

Topic	Count	OpenAI	KeyBERT	Representative Document
41	168	Shareholder Lawsuit Filings	lawsuit filed, antitrust lawsuit	Lawsuits Filed Against ATIP, SESN and APPH - Jakubowitz Law Pursues Shareholders Claims
42	161	5G and Wireless Network Collaboration	broadband, provider	Mediacom Communications taps Casa Systems' fixed wireless and packet core solutions to expand availability of high-speed broadband connectivity in hard-to-reach communities
43	161	EV Market Dynamics	tesla, dow jones	Dow Jones Futures: Jobs Report Key For Choppy Market; Tesla Cuts Model 3, Y Prices Again
44	157	Downgrades to Hold from Buy	downgraded hold buy, downgraded sell	Truist downgraded to hold from buy at Jefferies
45	146	New Credit Facility Announcements	credit facility, new credit	The ONE Group Announces Increase to Credit Facility Up to $87 Million
46	146	Water System Investments	water company, water technology	Virginia American Water Signs Agreement to Purchase Town of Cape Charles Water and Sewer Systems
47	134	Retail Store Openings	ross store, ross	Ross Dress for Less to Open a New Store in Concord, California
48	131	H.C. Wainwright Investment Presentations	global investment conference, wainwright global investment	Azitra to Present at the H.C. Wainwright 25th Annual Global Investment Conference
49	126	Regular Dividend Declarations	dividend declares, declares dividend	GM Declares Quarterly Dividend
50	125	Needham Annual Growth Conference	growth conference, virtual growth conference	Petco to Participate in the 25th Annual Needham Growth Conference
51	125	Russell Index Inclusions	russell, index	Nuvectis Pharma Set to Join the Russell 2000(R) and Russell 3000(R) Indexes
52	125	Insider Trading in Bancshares	commerce bancshares, capital bancshares	Dir Burkhead Buys 1000 Of Hawthorn Bancshares Inc
53	120	Fitness Equipment Revenue Decline	peloton, earnings	Officer Rendich Acquires 18788 Of Peloton Interactive Inc
54	111	Director Stock Transactions	bank dir, bancorporation	Dir Berta Buys 2000 Of First Financial Bancorp/OH
55	110	Technology & Telecom Conferences	telecom conference, medium telecom conference	TaskUs to Present at Morgan Stanley Technology, Media & Telecom Conference
56	106	Securities Fraud Class Actions	security fraud, fraud	Kessler Topaz Meltzer & Check, LLP Announces a Securities Fraud Class Action Filed Against Camber Energy, Inc. (CEI)
57	104	Recognized Corporate Excellence	company consecutive, henry schein	Brown-Forman Named One of the World's Most Ethical Companies in 2022 by Ethisphere
58	101	Early Black Friday Apple Deals	iphone, verizon	Black Friday iPhone Deals 2022: Top Early Apple iPhone 14, 13, 12, 11, SE, XR & More Sales Found by Deal Stripe
59	101	Food Company Sales Growth	good food company, real good food	Simply Good Foods Records 1Q Rev Growth

Table OA7: Explanations Topics

This table shows the topics generated using LLM explanations. The column "OpenAI" contains the label generated by ChatGPT. The column "KeyBERT" contains the representative words. The column "Representative Document" contains the document closest to the center of the cluster. Topic -1 corresponds to all the headlines not classified into clusters.

Topic	Count	OpenAI	KeyBERT	Representative Document
-1	40456	Baseline	NA	NA
0	10514	Conference Visibility for Investors	conference increase visibility, visibility attract investor	Participation in a healthcare conference can increase visibility and potentially attract investors.
1	6933	Executive Share Surrender Signal	price surrendering share, stock price surrendering	The surrendering of shares by a VP could indicate a lack of confidence in the company's future performance, which may negatively impact the stock price.
2	4770	New Executive Appointments	appointment new board, new board member	The appointment of a new board member can bring fresh perspectives and strategies, potentially boosting the company's performance and stock price.
3	3348	Positive Investor Dividends	dividend positive, cash dividend	Declaring a cash dividend is generally seen as a positive sign for investors, indicating the company's financial stability and commitment to shareholder returns.
4	1611	Executive Share Confidence Signals	lack confidence company, indicates lack confidence	The sale of a significant number of shares by a company executive may indicate a lack of confidence in the company's future performance.
5	1367	Record Results Boost Stock	performance positive stock, stock price record	The record results indicate strong financial performance, which is likely to boost the stock price in the short term.
6	1250	Strong Profit Impact on Stocks	price higher profit, stock price higher	Higher profit and sales indicate strong financial performance, which is likely to boost the stock price in the short term.
7	1035	Chairman Share Transactions	strong insider confidence, insider confidence	This indicates strong insider confidence in the company's future performance, which can positively influence the stock price.
8	1021	Cloud Partnerships Boost Revenue	increase alphabet revenue, alphabet revenue	This partnership could potentially enhance Alphabet's revenue through Google Cloud services.
9	934	CFO Stock Confidence	performance cfo investing, cfo investing	This indicates confidence in the company's future performance as the CFO is investing his own money in the company.
10	773	Renewable Energy Partnerships	environmentally conscious investor, investor partnership	This agreement demonstrates a commitment to clean energy and could attract environmentally conscious investors.
11	741	Insider Confidence in Company	indicates insider confidence, insider confidence	This indicates insider confidence in the company's future performance.
12	730	Stock Price Reaction to Results	stock price negatively, stock price positively	The stock price is likely to increase due to better-than-expected financial results.
13	699	Investor Conference Participation	participation investor conference, presenting investor conference	Participation in investor conferences can increase visibility and potentially attract new investors.
14	680	S&P Credit Rating Impacts	stock price rating, rating indicates stable	The rating 'BBB-' by S&PGR indicates a stable outlook for Clean Harbors, which is positive for the stock price.
15	552	EPS Consensus Exceed Boost	earnings share eps, earnings share exceeded	The company's earnings per share (EPS) exceeded the consensus estimate, which is typically positive for the stock price.
16	551	Stock Repurchase Confidence	indicates company confidence, indicates company confident	This indicates the company's confidence in its own stock, which can boost investor sentiment and potentially increase the stock price.
17	518	CEO Stock Purchase Influence	ceo acquisition significant, ceo acquisition additional	The CEO's acquisition of additional shares demonstrates confidence in the company's future, which can positively influence the stock price.
18	510	Merger Investigation Uncertainty	short term merger, stock price acquisition	The investigation of the merger could create uncertainty and negatively impact the stock price in the short term.
19	487	Capital Raising for Growth	indicates company raising, company raising capital	This indicates the company is raising capital, which can be used for growth and expansion.

Topic	Count	OpenAI	KeyBERT	Representative Document
20	485	Executive Share Gifting Confidence	gifting share confidence, share indicates confidence	A director gifting shares can be seen as a positive sign of confidence in the company's future performance.
21	481	Stock Price Target Outlook	stock price indicates, stock price target	The increase in the stock price target indicates a positive outlook for CarMax's stock in the short term.
22	418	Increased Aircraft Demand Impact	airline stock price, airline stock	This indicates increased demand and profitability for United Airlines, which is likely to boost its stock price in the short term.
23	329	Increased Investor Confidence Impact	indicates increased investor, stock price indicates	This indicates increased confidence in the company's value, which can positively influence the stock price.
24	323	Award Impact on Stocks	stock price award, recognition boost investor	This award recognition can boost investor confidence and positively impact Sientra's stock price in the short term.
25	319	Engaging Analysts for Stock Boost	stock price indicates, investor analyst lead	This indicates that HCI Group is actively engaging with investors and analysts, which could potentially boost its stock price.
26	309	Investor Engagement through Conferences	stock price indicates, investor indicates	This indicates that Flowserve is actively engaging with investors and analysts, which could potentially boost its stock price.
27	301	Stock Price Impact Forecast	likely decrease company, price likely decrease	The stock price is likely to decrease due to the negative forecast.
28	275	Earnings Conference Calls Impact	investor gain insight, announcement earnings release	The announcement of earnings release and conference call can lead to increased investor interest and potential stock price movement.
29	272	Fitch Ratings Affirmations	fitch indicates stable, confidence company stability	Fitch's affirmation of Navient Corporation's rating with a stable outlook indicates confidence in the company's financial stability.
30	245	Strong Performance Boosts Stocks	performance positive stock, positively impact stock	The increase in profit due to eased restrictions and increased travel indicates positive growth for MGM Resorts International's stock price in the short term.
31	241	Fitch Credit Rating Affirmations	fitch indicates positive, rating fitch indicates	Fitch's affirmation of Steel Dynamics' IDR at 'BBB' with a stable outlook indicates a positive credit rating, which is good for the stock price in the short term.
32	224	Securities Class Action Impact	stock price lawsuit, class action lawsuit	A securities class action lawsuit can negatively impact the stock price in the short term.
33	216	Law Firm Investigations Impact	stock price investigation, firm reveal negative	The investigation by a law firm could potentially lead to negative outcomes for Neenah's stock price in the short term.
34	201	Negative Impact of Reverse Stock Split	reverse stock split, reverse stock	A reverse stock split often indicates a company is in financial trouble, which could negatively impact the stock price.
35	195	Investor Conference Presentations	new investor presentation, stock price presentation	This presentation could potentially attract new investors and increase the stock price.
36	190	Partnership Impact on Revenue	stock price partnership, customer base partnership	This partnership could potentially increase Affirm's customer base and revenue, positively impacting its stock price.
37	188	Record Earnings Boost Stocks	price record earnings, performance positive stock	Record results for the quarter indicate strong financial performance, which is likely to boost the stock price in the short term.
38	176	Contract Awards Boosting Revenue	contract award increase, price contract award	This contract award could increase Astronics Corporation's revenue, which is likely to positively impact its stock price in the short term.
39	174	Positive Dividend Announcement	declaring cash dividend, cash dividend	Declaring a quarterly cash dividend is generally seen as a positive sign for investors, indicating the company's financial stability and commitment to shareholder returns.
40	170	Regulatory Fines Impacting Stocks	negatively impacting stock, microsoft stock price	Microsoft may face regulatory fines and restrictions, negatively impacting its stock price in the short term.

Topic	Count	OpenAI	KeyBERT	Representative Document
41	161	Strategic Partnerships Boosting Market Reach	partnership likely boost, market partnership	This collaboration will likely boost Qualcomm's presence in the growing private 5G network market, potentially increasing demand for their products and services.
42	161	Stock Price Fluctuations	tesla stock, stock price negatively	Elon Musk's comments causing Tesla's stock to skid late could negatively impact the stock price in the short term.
43	159	Lawsuit Impact on Stocks	stock price lawsuit, negatively impacting stock	Lawsuits against a company can negatively impact its stock price in the short term.
44	151	Stock Downgrade Analysis	stock price downgrade, downgrade suggests analyst	The downgrade suggests that the analyst believes the stock may not provide a good return in the near future.
45	148	Increased Credit Flexibility	credit facility provides, credit facility	The new credit facility provides Ensign Group with increased financial flexibility and resources for growth.
46	146	Fitch Credit Ratings	fitch rating indicates, rating fitch indicates	Fitch's rating of 'BBB-' indicates a stable investment grade, which is positive for PG&E's stock price in the short term.
47	145	New Store Openings Impact	stock price opening, new store indicates	The opening of a new store indicates business expansion, which is typically a positive sign for stock prices.
48	145	Water Utilities Growth	increased revenue, acquisition increase	This acquisition could potentially increase American Water Works Co.'s revenue and customer base.
49	140	New Store Expansion Benefits	new store indicates, new store	The opening of a new store indicates expansion and potential growth in revenue for Ross Stores.
50	129	Global Investment Conference Participation	global investment conference, investment conference attract	Participation in a global investment conference can increase visibility and potentially attract new investors.
51	126	Positive Dividend Declaration	dividend positive sign, dividend positive	Declaring a quarterly dividend is generally seen as a positive sign for investors, indicating the company's financial stability and commitment to shareholder returns.
52	126	Investor Engagement at Conferences	investor participation conference, conference increase visibility	Participation in a growth conference can increase visibility and potentially attract new investors.
53	122	Insider Confidence Indicators	indicates insider confidence, insider confidence	This indicates insider confidence in the company's future performance.
54	108	Insider Trading Signals	indicates insider confidence, insider confidence	This indicates insider confidence in the company's future performance.
55	108	Securities Fraud Lawsuit Impact	class action lawsuit, lawsuit negatively impact	A securities fraud class action lawsuit can negatively impact the stock price in the short term.
56	108	Investor Engagement Events	stock price indicates, investor analyst boost	This indicates that the company is actively engaging with investors and analysts, which could potentially boost its stock price.
57	103	Black Friday Sales Impact	apple stock price, apple stock	Increased sales due to Black Friday deals can positively impact Apple's stock price in the short term.
58	101	Stock Price Impact	performance positive stock, positive stock	The decrease in profit and tightening margins could negatively impact Simply Good Foods' stock price in the short term.
59	100	Stock Impact from Corporate Actions	positively impact stock, negatively impact stock	This indicates potential growth and expansion for Blackstone, which could positively impact their stock price.

Appendix B: Model Summaries

In this section, we present an overview of the ten different models that we study in this paper. We order them by their release date.

Model 1. GPT-1: Estimated Number of Parameters: 117 million, Release Date: Feb 2018, Website: https://huggingface.co/docs/transformers/model_doc/openai-gpt.

Generative Pre-trained Transformer 1 (GPT-1) was the first of OpenAI's large language models following Google's invention of the transformer architecture in 2017. It was introduced in February 2018 by OpenAI. GPT-1 had 117 million parameters and significantly improved previous state-of-the-art language models. One of its strengths was its ability to generate fluent and coherent language when given a prompt or context. It was based on the transformer architecture and trained on a large corpus of books.

Model 2. BERT: Estimated Number of Parameters: 110 million, Release Date: Nov 2, 2018, Website: https://huggingface.co/bert-base-uncased.

BERT (Bidirectional Encoder Representations from Transformers) is a family of language models introduced in 2018 by researchers at Google. It is based on the transformer architecture and was initially implemented in English at two model sizes: BERT BASE and BERT Large. Both models were pre-trained on the Toronto BookCorpus and English Wikipedia. BERT was pre-trained simultaneously on language modeling and next-sentence prediction. As a result of this training process, BERT learns latent representations of words and sentences in context. It can be fine-tuned with fewer resources on smaller datasets to optimize its performance on specific tasks such as NLP tasks and sequence-to-sequence-based language generation tasks.

Model 3. BERT-Large: Estimated Number of Parameters: 336 million, Release Date: Nov 2, 2018, Website: https://huggingface.co/bert-large-uncased.

BERT (Bidirectional Encoder Representations from Transformers) is a family of language models introduced in 2018 by researchers at Google. It is based on the transformer architecture and was initially implemented in English at two model sizes: BERT BASE and BERT Large. Both models were pre-trained on the Toronto BookCorpus and English Wikipedia. BERT was pre-trained simultaneously on language modeling and next-sentence prediction. As a result of this training process, BERT learns latent representations of words and sentences in context. It can be fine-tuned with fewer resources on smaller datasets to optimize its performance on specific tasks such as NLP tasks and sequence-to-sequence-based

language generation tasks.

Model 4. GPT-2: Estimated Number of Parameters: 124 million, Release Date: Feb 14, 2019, Website: https://huggingface.co/gpt2.

Generative Pre-trained Transformer 2 (GPT-2) is a large language model by OpenAI, the second in their foundational series of GPT models1. It was pre-trained on BookCorpus, a dataset of over 7,000 unpublished fiction books from various genres, and trained on a dataset of 8 million web pages1. GPT-2 was partially released in February 2019. It is a decoder-only transformer model of deep neural networks, which uses attention in place of previous recurrence- and convolution-based architectures. The model demonstrated strong zero-shot and few-shot learning on many tasks. This is the smallest version of GPT-2, with 124M parameters.

Model 5. GPT-2-Large: Estimated Number of Parameters: 774 million, Release Date: Feb 1, 2019, Website: https://huggingface.co/gpt2-large.

GPT-2 Large is the 774M parameter version of GPT-2. Generative Pre-trained Transformer 2 (GPT-2) is a large language model by OpenAI, the second in their foundational series of GPT models1. It was pre-trained on BookCorpus, a dataset of over 7,000 unpublished fiction books from various genres, and trained on a dataset of 8 million web pages1. GPT-2 was partially released in February 2019. It is a decoder-only transformer model of deep neural networks, which uses attention in place of previous recurrence- and convolution-based architectures. The model demonstrated strong zero-shot and few-shot learning on many tasks.

Model 6. BART-Large: Estimated Number of Parameters: 400 million, Release Date: Oct 29, 2019, Website: https://huggingface.co/facebook/bart-large-mnli.

BART (large-sized model) is a pre-trained model on the English language, introduced in the paper "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension" by Lewis et al. (2019). It uses a standard seq2seq/machine translation architecture with a bidirectional encoder (like BERT) and a left-to-right decoder (similar to GPT). The pre-training task involves randomly shuffling the order of the original sentences and a novel in-filling scheme, where text spans are replaced with a single mask token. BART is particularly effective when fine-tuned for text generation but also works well for comprehension tasks. It matches the performance of RoBERTa with comparable training resources on GLUE and SQuAD. It achieves new state-of-the-art results on a range of abstractive dialogue, question-answering, and summarization tasks, with gains

of up to 6 ROUGE. BART (large-sized model) has nearly 400M parameters.

Model 7. DistilBart-MNLI-12-1: Estimated Number of Parameters: $<$ 400 million , Release Date: Sep 21, 2020, Website: https://huggingface.co/valhalla/distilbart-mnli-12-1.

Distilbart-Mnli-12-1 is a distilled version of bart-large-mnli created using the No Teacher Distillation technique proposed for BART summarisation by Huggingface. It was released on September 21, 2020. It copies alternating layers from bart-large-mnli and is fine-tuned more on the same data. The performance drop is minimal compared to the original model.

Model 8. GPT-3.5: Estimated Number of Parameters: 175 billion, Release Date: Nov 30, 2022, Website: https://platform.openai.com/docs/models.

GPT-3.5 is a fined-tuned version of the GPT-3 (Generative Pre-Trained Transformer) model. It has 175 billion parameters and is trained on a dataset of text and code up to June 2021. GPT-3.5 models can understand and generate natural language or code. The most capable and cost-effective model in the GPT-3.5 family is gpt-3.5-turbo, which has been optimized for chat using the Chat completions API but works well for traditional completions tasks. GPT-3.5 effectively performs various tasks, including text generation, translation, summarization, question answering, code generation, and creative writing.

Model 9. GPT-4 or ChatGPT 4: Estimated Number of Parameters: 1.76 trillion, Release Date: Mar 14, 2023, Website: https://platform.openai.com/docs/models

GPT-4 is a multimodal large language model created by OpenAI and the fourth in its series of GPT foundation models. OpenAI released it on March 14, 2023. As a transformer-based model, GPT-4 uses a paradigm where pre-training using both public data and "data licensed from third-party providers" is used to predict the next token. After this step, the model was fine-tuned with reinforcement learning feedback from humans and AI for human alignment and policy compliance. OpenAI did not release the technical details of GPT-4; the technical report explicitly refrained from specifying the model size, architecture, or hardware used during either training or inference. GPT-4 has several capabilities, including generating text that is indistinguishable from human-written text; translating languages with high accuracy; writing different kinds of creative content, such as poems, code, scripts, musical pieces, emails, and letters; and answering questions in an informative way, even if they are open-ended, challenging, or strange.

Model 10. RavenPack Estimated Number of Parameters: NA, Release Date: NA, Website: https://www.ravenpack.com/

RavenPack is a leading provider of news analytics data. For each news item in their database, RavenPack Analytics generates an event sentiment score using their proprietary algorithms.

Model 11. FinBERT Estimated Number of Parameters: 774 million, Release Date: 27 Aug 2019, Website: https://huggingface.co/ProsusAI/finbert

FinBERT, introduced by Dogu Araci, is a pre-trained NLP model fine-tuned for financial sentiment classification. It leverages the BERT language model, further trained on a large financial corpus, making it effective for sentiment analysis tasks in the financial domain. The model, which relies on Hugging Face's pytorch_pretrained_bert library, is available on Hugging Face's model hub and their GitHub repository.

Appendix C: Prompts for Other LLMs

This appendix provides details on the prompts for other LLMs. While a critical focus of our paper is on ChatGPT, we compare the results of ChatGPT with those of more basic models such as BERT, GPT-1, and GPT-2. We employ a different strategy because those models cannot follow instructions or answer specific questions.

GPT-1 and GPT-2 are autocomplete models. Hence, we use the following sentence that the models complete:

> News: + headline + f'Will this increase or decrease the stock price of firm? This will make firm's stock price go "

The usual response is "up," "down," followed by a brief sentence fragment. The answers are usually not fully legible but include positive and negative words. We count the positive words against the negative words and assign a $+1$ for every positive and a -1 for every negative. We then consider the sentiment positive if the sum is positive and vice versa. The positive words are 'up,' 'high,' 'sky,' 'top,' 'increase,' 'stratosphere,' 'boom,' 'roof,' 'skyrocket,' 'soar,' 'surge,' 'climb,' 'rise,' 'rising,' 'expand,' 'flourish.' The negative words are 'down,' 'low,' 'bottom,' 'decrease,' 'back,' 'under,' 'plummet,' 'drop,' 'decline,' 'tumble,' 'fall,' 'contract,' 'struggle.'

BERT is only able to complete one word out of a sentence. Hence, we ask it to complete the following sentence:

> Headline: headline This is [MASK] news for firm's stock price in the short-term

Where [MASK] is the corresponding word that BERT will input. The answers set consists of 'good,' 'the,' 'big,' and 'bad.' We classify 'good' as +1, 'bad' as -1, and the others as zero.

The BART model is capable of zero-shot classification. This means it can classify text according to predefined categories without seeing examples of what corresponds to a good category. We provide each headline and then classify it into one of the following categories:

1. good news for the stock price of firm in the short term
2. bad news for the stock price of firm in the short term
3. not news for the stock price of firm in the short term

We then assign a numerical score of +1 for good, -1 for bad, and 0 for not.

Appendix D: Model Proofs

In this section, we provide the proofs for the theoretical results in Section 2 of the paper. We use Mathematica 14.0 to verify all the sign inequalities, derivatives, and expectations.

A Attentive Agent's Demand

Substituting the attentive agent's optimal demand, the value function can be written as:

$$V_2(w_2) = \max_{x_3} -exp\{-\alpha E[\tilde{w_3}] + \frac{\alpha^2}{2} Var(\tilde{w_3})\} \tag{37}$$

$$= \max_{x_3} -exp\{-\alpha(w_2 + x_3(\mu_{A|s} - p_2)) + \frac{\alpha^2}{2} x^2 \sigma^2_{d,A|s}\} \tag{38}$$

$$= -exp\{-\alpha(w_2 + \frac{\mu_{A|s} - p_2}{\alpha\sigma^2_{d,A|s}}(\mu_{A|s} - p_2)) + \frac{\alpha^2}{2}(\frac{\mu_{A|s} - p_2}{\alpha\sigma^2_{d,A|s}})^2 \sigma^2_{d,A|s}\} \tag{39}$$

$$= -exp\{-\alpha(w_2) - \frac{(\mu_{A|s} - p_2)^2}{\sigma^2_{d,A|s}} + \frac{1}{2}\frac{(\mu_{A|s} - p_2)^2}{\sigma^2_{d,A|s}}\} \tag{40}$$

$$= -exp\{-\alpha(w_2) - \frac{1}{2}\frac{(\mu_{A|s} - p_2)^2}{\sigma^2_{d,A|s}}\}. \tag{41}$$

In period 1, attentive agents maximize the expected period 2 value function:

$$V1(w_1) = \max_{x_2} E_1[V2(\tilde{w_2})] \tag{42}$$

$$\tilde{w_2} = w_1 + x_2(\tilde{p_2} - p_1), \tag{43}$$

which we can rewrite as

$$V1(w_1) = \max_x E_1[V_2(\tilde{w_2})] \tag{44}$$

$$= \max_x -E_1[exp\{-\alpha(w_2) - \frac{1}{2}\frac{(\mu_{A|s} - \tilde{p_2})^2}{\sigma^2_{d,A|s}}\}] \tag{45}$$

$$= \max_x -E_1[exp\{-\alpha(w_1 + x(\tilde{p_2} - p_1)) - \frac{1}{2}\frac{(\mu_{A|s} - \tilde{p_2})^2}{\sigma^2_{d,A|s}}\}] \tag{46}$$

$$= \max_x -E_1[exp\{a(x) + b(x)\tilde{p_2} + c\tilde{p_2}^2\}] \tag{47}$$

with $a = a(x) = \alpha p_1 x - \frac{\mu^2_{A|s}}{2\sigma^2_{d,A|s}} - \alpha w_1$, $b = b(x) = \frac{\mu_{A|s}}{\sigma^2_{d,A|s}} - \alpha x$, and $c = -\frac{1}{2\sigma^2_{d,A|s}}$.

To get a closed-form solution, we use the following Lemma:

Lemma A.1. *Let $z \sim N(\mu, \sigma^2)$. Then, if $1 - 2c\sigma^2 > 0$,*

$$E[\exp(a + bz + cz^2)] = \frac{\exp\left(-\frac{\sigma^2\left(b^2 - 4ac\right) + 2(a + \mu(b + c\mu))}{4c\sigma^2 - 2}\right)}{\sqrt{1 - 2c\sigma^2}}. \tag{48}$$

Proof. The omitted proof follows from completing the square in the probability density function. □

$$G(w, x) = \frac{\exp\left(\frac{-4ac\sigma_p^2 + 2a + b^2\sigma_p^2 + 2b\mu_p + 2c\mu_p^2}{2 - 4c\sigma_p^2}\right)}{\sqrt{1 - 2c\sigma_p^2}}. \tag{49}$$

Substituting the values of a, b, and c, we obtain:

$$G(w,x) \equiv -\frac{\exp\left(-\frac{2\left(\alpha p_1 x-\frac{\mu_{A|s}^2}{2\sigma_{d,A|s}^2}+\mu_{A|s}\left(\frac{\mu_{A|s}}{2\sigma_{d,A|s}^2}-\alpha x\right)-\alpha w\right)+\alpha^2\sigma_{A,s}^4\sigma_u^2\left(\frac{2\left(\alpha p_1 x-\frac{\mu_{A|s}^2}{2\sigma_{d,A|s}^2}-\alpha w\right)}{\sigma_{d,A|s}^2}+\left(\frac{\mu_{A|s}}{\sigma_{d,A|s}^2}-\alpha x\right)^2\right)}{-2\alpha^2\sigma_{d,A|s}^2\sigma_u^2-2}\right)}{\sqrt{\alpha^2\sigma_{d,A|s}^2\sigma_u^2+1}}. \tag{50}$$

We then use the first-order condition:

$$\frac{\partial G}{\partial x}|_{x=x_{2,A}} = -\frac{\left(2(\alpha p_1-\alpha\mu_{A|s})+\alpha^2\sigma_{d,A|s}^4\sigma_u^2\left(\frac{2\alpha p_1}{\sigma_{d,A|s}^2}-2\alpha\left(\frac{\mu_{A|s}}{\sigma_{d,A|s}^2}-\alpha x_{2,A}\right)\right)\right)G(w_1,x_{2,A})}{\left(-2\alpha^2\sigma_{d,A|s}^2\sigma_u^2-2\right)\sqrt{\alpha^2\sigma_{d,A|s}^2\sigma_u^2+1}} = 0 \tag{51}$$

$$\Leftrightarrow \tag{52}$$

$$2(\alpha p_1-\alpha\mu_{A|s})+\alpha^2\sigma_{d,A|s}^4\sigma_u^2\left(\frac{2\alpha p_1}{\sigma_{d,A|s}^2}-2\alpha\left(\frac{\mu_{A|s}}{\sigma_{d,A|s}^2}-\alpha x_{2,A}\right)\right) = 0 \tag{53}$$

$$\Rightarrow \tag{54}$$

$$x_{2,A} = \frac{(\mu_{A|s}-p_1)\left(\alpha^2\sigma_{d,A|s}^2\sigma_u^2+1\right)}{\alpha^3\sigma_{d,A|s}^4\sigma_u^2} \tag{55}$$

Notice that we can write the demand function as

$$x_{2,A} = \frac{(\mu_{A|s} - p_1)\left(\alpha^2\sigma^2_{d,A|s}\sigma^2_u + 1\right)}{\alpha^3\sigma^4_{d,A|s}\sigma^2_u} \tag{56}$$

$$= \frac{(\mu_{A|s} - p_1)\left(\alpha^2\sigma^2_{d,A|s}\sigma^2_u\right)}{\alpha^3\sigma^4_{d,A|s}\sigma^2_u} + \frac{(\mu_{A|s} - p_1)}{\alpha^3\sigma^4_{d,A|s}\sigma^2_u} \tag{57}$$

$$= \frac{(\mu_{A|s} - p_1)}{\alpha\sigma^2_{d,A|s}} + \frac{(\mu_p - p_1)}{\alpha\sigma^2_p}. \tag{58}$$

B Proposition 2.1 - Mispricing

Mispricing is defined as:

$$E[\alpha^2_M] \equiv E[(E[\tilde{d}|s] - E[p_1|s])^2]. \tag{59}$$

Evaluating the expectation, we get:

$$E[\alpha^2_M] = \tag{60}$$

$$\frac{\tau_D^2\tau_S\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(-\pi_A\gamma_A-\left(\left(\left(-\pi_I\right)\right)\tau_D\sigma_\xi^2\right)+\left(\left(-\pi_I\right)\right)\gamma_A\left(\omega\tau_D\sigma_\xi^2-\sigma_\xi^2\tau_S+\omega\right)+\left(\left(-\pi_I\right)\right)\omega\gamma_A^2\sigma_\xi^2\tau_S+1\right)-\pi_A V^2\left(\gamma_A-1\right)\left(\tau_A+\tau_D\right)\right)^2}{\left(\tau_D+\tau_S\right)^2\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\gamma_A\tau_S+\tau_D\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\tau_D^2\sigma_\xi^2+\tau_D\left(\left(\left(-\pi_I\right)\right)\left(\omega+1\right)\gamma_A\sigma_\xi^2\tau_S-1\right)+\tau_A\left(\left(\left(-\pi_I\right)\right)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)-\pi_A\right)\right)-\pi_A V^2\left(\left(\tau_{\mu,A|s}\right)\right)^2\right)^2} \tag{61}$$

The derivative w.r.t. the quantity of attentive agents, π_A, is:

$$\frac{\partial E[\alpha_M^2]}{\partial \pi_A} = \tag{62}$$

$$-\left(2\alpha^2(\omega-1)\gamma_A\tau_D^2\tau_S\sigma_u^2\times\right. \tag{63}$$

$$\left(\sigma_\xi^2\left(\tau_A+\tau_D\right)+1\right)^2\times \tag{64}$$

$$\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)+V^2\left(\left(\tau_{\mu,A|s}\right)\right)\right)\times \tag{65}$$

$$\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\tau_A+\tau_D\right)+1\right)\left(-\pi_A\gamma_A-\left(\left(\left(-\pi_I\right)\right)\tau_D\sigma_\xi^2\right)+\left(\left(-\pi_I\right)\right)\gamma_A\left(\omega\tau_D\sigma_\xi^2-\sigma_\xi^2\tau_S+\omega\right)+\left(\left(-\pi_I\right)\right)\omega\gamma_A^2\sigma_\xi^2\tau_S+1\right)-\pi_AV^2\left(\gamma_A-1\right)\left(\tau_A+\tau_D\right)\right)/ \tag{66}$$

$$\left(\left(\tau_D+\tau_S\right)\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\gamma_A\tau_S+\tau_D\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\tau_D^2\sigma_\xi^2+\tau_D\left(\left(\left(-\pi_I\right)\right)\left(\omega+1\right)\gamma_A\sigma_\xi^2\tau_S-1\right)+\tau_A\left(\left(\left(-\pi_I\right)\right)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)-\pi_A\right)\right)-\pi_AV^2\left(\left(\tau_{\mu,A|s}\right)\right)^2\right)^3\right). \tag{67}$$

This derivative is negative since every term in the numerator is negative, and the denominator is positive.
The derivative w.r.t. the information capacity of attentive agents, γ_A, is:

$$\frac{\partial E[\alpha_M^2]}{\partial \gamma_A} = \tag{68}$$

$$\left(2\tau_D^2\tau_S\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(-\pi_A\gamma_A-\left(\left(\left(-\pi_I\right)\right)\tau_D\sigma_\xi^2\right)+\left(\left(-\pi_I\right)\right)\gamma_A\left(\omega\tau_D\sigma_\xi^2-\sigma_\xi^2\tau_S+\omega\right)+\left(\left(-\pi_I\right)\right)\omega\gamma_A^2\sigma_\xi^2\tau_S+1\right)-\pi_AV^2\left(\gamma_A-1\right)\left(\left(\tau_{\mu,A|s}\right)\right)\right)\times \tag{69}$$

$$\left(\alpha^4\sigma_u^4\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)^2\left(\left(\pi_A-1\right)\left(\omega\left(\left(\left(-\pi_I\right)\right)\sigma_\xi^4\left(\left(\tau_{\mu,A|s}\right)\right)^2+\sigma_\xi^2\left(\left(\pi_A-2\right)\tau_D-2\tau_A\right)-1\right)-\pi_A\tau_D\sigma_\xi^2\right)+\pi_A\right)- \tag{70}$$

$$\alpha^2\pi_AV^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\left(\omega+1\right)\tau_D^2\sigma_\xi^2+\tau_D\left(\left(\left(-\pi_I\right)\right)\left(3\omega+1\right)\gamma_A\sigma_\xi^2\tau_S+\pi_A\omega-\pi_A-\omega-1\right)+2\tau_A\left(\left(\left(-\pi_I\right)\right)\omega\gamma_A\sigma_\xi^2\tau_S-1\right)\right)+\pi_A^2V^4\left(\left(\tau_{\mu,A|s}\right)\right)^2\right)/ \tag{71}$$

$$\left(\left(\tau_D+\tau_S\right)\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\tau_D^2\sigma_\xi^2+\tau_D\left(\left(\left(-\pi_I\right)\right)\left(\omega+1\right)\gamma_A\sigma_\xi^2\tau_S-1\right)+\tau_A\left(\left(\left(-\pi_I\right)\right)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)-\pi_A\right)\right)-\pi_AV^2\left(\left(\tau_{\mu,A|s}\right)\right)^2\right)^3\right). \tag{72}$$

This derivative is negative since every term in the numerator is negative, and the denominator is positive.
The derivative w.r.t. the information capacity of inattentive agents, ω, is:

$$\frac{\partial E[\alpha_M^2]}{\partial \omega} = \tag{73}$$

$$\left(2\alpha^2\left(\left(-\pi_I\right)\right)\gamma_A\tau_D^2\tau_S\sigma_u^2\left(\sigma_\xi^2\left(\gamma_A\tau_S+\tau_D\right)+1\right)^2\times\right. \tag{74}$$

$$\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\tau_A+\tau_D\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)-1\right)-\pi_AV^2\left(\left(\tau_{\mu,A|s}\right)\right)\right)\times \tag{75}$$

$$\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\tau_A+\tau_D\right)+1\right)\left(-\pi_A\gamma_A-\left(\left(\left(-\pi_I\right)\right)\tau_D\sigma_\xi^2\right)+\left(\left(-\pi_I\right)\right)\gamma_A\left(\omega\tau_D\sigma_\xi^2-\sigma_\xi^2\tau_S+\omega\right)+\left(\left(-\pi_I\right)\right)\omega\gamma_A^2\sigma_\xi^2\tau_S+1\right)-\pi_AV^2\left(\gamma_A-1\right)\left(\tau_A+\tau_D\right)\right)/ \tag{76}$$

$$\left(\left(\tau_D+\tau_S\right)\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\gamma_A\tau_S+\tau_D\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\tau_D^2\sigma_\xi^2+\tau_D\left(\left(\left(-\pi_I\right)\right)\left(\omega+1\right)\gamma_A\sigma_\xi^2\tau_S-1\right)+\tau_A\left(\left(\left(-\pi_I\right)\right)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)-\pi_A\right)\right)-\pi_AV^2\left(\left(\tau_{\mu,A|s}\right)\right)^2\right)^3\right). \tag{77}$$

This derivative is negative since the last term in the numerator is negative, while all other terms are positive.
The derivative w.r.t. the total volume of traders, V, is also negative and given by:

$$\frac{\partial E[\alpha_M^2]}{\partial V} = \tag{78}$$

$$(4\alpha^2\left((-\pi_I)\right)\pi_A V(\omega-1)\gamma_A\tau_D^2\tau_S\sigma_u^2\left(\tau_A+\tau_D\right)\times \tag{79}$$

$$\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)^2\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(-\pi_A\gamma_A-\left(\left((-\pi_I)\right)\tau_D\sigma_\xi^2\right)+\left((-\pi_I)\right)\gamma_A\left(\omega\tau_D\sigma_\xi^2-\sigma_\xi^2\tau_S+\omega\right)+\left((-\pi_I)\right)\omega\gamma_A^2\sigma_\xi^2\tau_S+1\right)-\pi_A V^2\left(\gamma_A-1\right)\left(\left(\tau_{\mu,A|s}\right)\right)\right)/ \tag{80}$$

$$\left(\left(\tau_D+\tau_S\right)\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(\left((-\pi_I)\right)\tau_D^2\sigma_\xi^2+\tau_D\left(\left(\pi_A-1\right)\left(\omega+1\right)\gamma_A\sigma_\xi^2\tau_S-1\right)+\tau_A\left(\left((-\pi_I)\right)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)-\pi_A\right)\right)-\pi_A V^2\left(\tau_A+\tau_D\right)^2\right)^3. \tag{81}$$

The derivative w.r.t. the risk aversion, α, is:

$$\frac{\partial E[\alpha_M^2]}{\partial \alpha} = \tag{82}$$

$$-\left(4\alpha\left((-\pi_I)\right)\pi_A V^2(\omega-1)\gamma_A\tau_D^2\tau_S\sigma_u^2\left(\tau_A+\tau_D\right)\times\right. \tag{83}$$

$$\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)^2\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(-\pi_A\gamma_A-\left(\left((-\pi_I)\right)\tau_D\sigma_\xi^2\right)+\left((-\pi_I)\right)\gamma_A\left(\omega\tau_D\sigma_\xi^2-\sigma_\xi^2\tau_S+\omega\right)+\left((-\pi_I)\right)\omega\gamma_A^2\sigma_\xi^2\tau_S+1\right)-\pi_A V^2\left(\gamma_A-1\right)\left(\left(\tau_{\mu,A|s}\right)\right)\right)/ \tag{84}$$

$$\left.\left(\left(\tau_D+\tau_S\right)\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(\left((-\pi_I)\right)\tau_D^2\sigma_\xi^2+\tau_D\left(\left(\pi_A-1\right)\left(\omega+1\right)\gamma_A\sigma_\xi^2\tau_S-1\right)+\tau_A\left(\left((-\pi_I)\right)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)-\pi_A\right)\right)-\pi_A V^2\left(\tau_A+\tau_D\right)^2\right)^3\right). \tag{85}$$

This derivative is positive since every term in the numerator is negative, and the denominator is positive.
The derivative w.r.t. the noise trader volatility, σ_u, is:

$$\frac{\partial E[\alpha_M^2]}{\partial \sigma_u} = \tag{86}$$

$$-\left(4\alpha^2\left((-\pi_I)\right)\pi_A V^2(\omega-1)\gamma_A\tau_D^2\tau_S\sigma_u\left(\tau_A+\tau_D\right)\times\right. \tag{87}$$

$$\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)^2\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(-\pi_A\gamma_A-\left(\left((-\pi_I)\right)\tau_D\sigma_\xi^2\right)+\left((-\pi_I)\right)\gamma_A\left(\omega\tau_D\sigma_\xi^2-\sigma_\xi^2\tau_S+\omega\right)+\left((-\pi_I)\right)\omega\gamma_A^2\sigma_\xi^2\tau_S+1\right)-\pi_A V^2\left(\gamma_A-1\right)\left(\left(\tau_{\mu,A|s}\right)\right)\right)/ \tag{88}$$

$$\left.\left(\left(\tau_D+\tau_S\right)\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(\left((-\pi_I)\right)\tau_D^2\sigma_\xi^2+\tau_D\left(\left(\pi_A-1\right)\left(\omega+1\right)\gamma_A\sigma_\xi^2\tau_S-1\right)+\tau_A\left(\left((-\pi_I)\right)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)-\pi_A\right)\right)-\pi_A V^2\left(\tau_A+\tau_D\right)^2\right)^3\right). \tag{89}$$

This derivative is positive.

C Theorem 1

The unconditional expectation of the profits using the LLM's signal is given by

$$\text{Profits}_{LLM} = E[x_{2,L}(\mu_{A|s} - p_1)] = \tag{90}$$

$$(\alpha\left((-\pi_I)\right)(\omega-1)\gamma_A\tau_D^2\tau_S\sigma_u^2\left(\sigma_\xi^2\left(\gamma_A\tau_S+\tau_D\right)+1\right)^2\times \tag{91}$$

$$(\lambda(c,k)\left(\pi_A V^2\left((\tau_{\mu,A|s})\right)-\alpha^2\sigma_u^2\left(\sigma_\xi^2\left((\tau_{\mu,A|s})\right)+1\right)\left(\left((-\pi_I)\right)\sigma_\xi^2\left(\gamma_A\tau_S+\tau_D\right)-1\right)\right)+ \tag{92}$$

$$\gamma_A\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\gamma_A\tau_S+\tau_D\right)+1\right)\left(\left((-\pi_I)\right)\omega\left(\sigma_\xi^2\left(\tau_A+\tau_D\right)+1\right)-\pi_A\right)-\pi_A V^2\left((\tau_{\mu,A|s})\right)\right))/ \tag{93}$$

$$\left(\left(\tau_A+\tau_D\right)\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left((\tau_{\mu,A|s})\right)+1\right)\left(\left(\pi_A-1\right)\tau_D^2\sigma_\xi^2+\tau_D\left(\left((-\pi_I)\right)\left(\omega+1\right)\gamma_A\sigma_\xi^2\tau_S-1\right)+\tau_A\left(\left((-\pi_I)\right)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)-\pi_A\right)\right)-\pi_A V^2\left((\tau_{\mu,A|s})\right)^2\right)^2\right). \tag{94}$$

To see that the profits increase with LLM model size, the derivative of the profits with respect to the model size is given by

$$\frac{\partial \text{Profits}_{LLM}}{\partial k} = \tag{95}$$

$$\frac{\alpha\left((-\pi_I)\right)(\omega-1)\gamma_A\tau_D^2\tau_S\sigma_u^2\frac{\partial\lambda}{\partial k}\left(\sigma_\xi^2\left((\tau_{\mu,A|s})\right)+1\right)^2\left(\pi_A V^2\left((\tau_{\mu,A|s})\right)-\alpha^2\sigma_u^2\left(\sigma_\xi^2\left((\tau_{\mu,A|s})\right)+1\right)\left(\left((-\pi_I)\right)\sigma_\xi^2\left(\gamma_A\tau_S+\tau_D\right)-1\right)\right)}{\left((\tau_{\mu,A|s})\right)\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left((\tau_{\mu,A|s})\right)+1\right)\left(\left((-\pi_I)\right)\tau_D^2\sigma_\xi^2+\tau_D\left(\left(\pi_A-1\right)\left(\omega+1\right)\gamma_A\sigma_\xi^2\tau_S-1\right)+\tau_A\left(\left((-\pi_I)\right)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)-\pi_A\right)\right)-\pi_A V^2\left(\tau_A+\tau_D\right)^2\right)^2}, \tag{96}$$

where both the numerator and the denominator are positive.

Moreover, the profits are positive if and only if

$$\lambda(c,k) > \text{Threshold}_\lambda \tag{97}$$

$$\equiv \frac{\gamma_A\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left((\tau_{\mu,A|s})\right)+1\right)\left(\left(\pi_A-1\right)\omega\left(\sigma_\xi^2\left((\tau_{\mu,A|s})\right)+1\right)-\pi_A\right)-\pi_A V^2\left(\gamma_A\tau_S+\tau_D\right)\right)}{\alpha^2\sigma_u^2\left(\sigma_\xi^2\left((\tau_{\mu,A|s})\right)+1\right)\left(\left((-\pi_I)\right)\sigma_\xi^2\left((\tau_{\mu,A|s})\right)-1\right)-\pi_A V^2\left(\gamma_A\tau_S+\tau_D\right)}. \tag{98}$$

For a fixed news complexity level c, let $\lambda_c(k) = \lambda(c,k)$. $\lambda_c(k)$ is a strictly increasing function of model size k and hence it is invertible. Thus, if we define

$$k^* \equiv \lambda_c^{-1}(\text{Threshold}_\lambda), \tag{99}$$

the profits of the LLM strategy are positive if and only if $k > k^*$. Since the inverse function is strictly increasing, it suffices to focus on the properties of Threshold_λ to characterize its behavior.

C.1 Threshold Properties

For inattentive agents' information capacity ω, the derivative of the threshold is positive and given by

$$\frac{\partial \text{Threshold}_\lambda}{\partial \omega} = \tag{100}$$

$$\frac{\alpha^2\left(\left(-\pi_I\right)\right)\gamma_A\sigma_u^2\left(\sigma_\xi^2\left(\tau_A+\tau_D\right)+1\right)^2}{\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)-1\right)-\pi_A V^2\left(\gamma_A\tau_S+\tau_D\right)}. \tag{101}$$

For attentive agents' information capacity γ_A, the derivative of the threshold is also positive and given by

$$\frac{\partial \text{Threshold}_\lambda}{\partial \gamma_A} = \tag{102}$$

$$\left(\alpha^4\sigma_u^4\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)^2\times\right. \tag{103}$$

$$\left(\left(\left(-\pi_I\right)\right)\left(\omega\left(\left(\left(-\pi_I\right)\right)\sigma_\xi^4\left(\left(\tau_{\mu,A|s}\right)\right)^2+\sigma_\xi^2\left(\left(\pi_A-2\right)\tau_D-2\tau_A\right)-1\right)-\pi_A\tau_D\sigma_\xi^2\right)+\pi_A\right)- \tag{104}$$

$$\alpha^2\pi_A V^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\times \tag{105}$$

$$\left(\left(\pi_A-1\right)\left(\omega+1\right)\tau_D^2\sigma_\xi^2+\tau_D\left(\left(\left(-\pi_I\right)\right)\left(3\omega+1\right)\gamma_A\sigma_\xi^2\tau_S+\pi_A\omega-\pi_A-\omega-1\right)+2\tau_A\left(\left(\left(-\pi_I\right)\right)\omega\gamma_A\sigma_\xi^2\tau_S-1\right)\right)+\pi_A^2V^4\left(\left(\tau_{\mu,A|s}\right)\right)^2)/ \tag{106}$$

$$\left.\left(\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\sigma_\xi^2\left(\tau_A+\tau_D\right)-1\right)-\pi_A V^2\left(\left(\tau_{\mu,A|s}\right)\right)\right)^2\right). \tag{107}$$

Moreover, for the proportion of attentive agents, π_A, the derivative of the threshold is also positive and given by

$$\frac{\partial \text{Threshold}_\lambda}{\partial \pi_A} = \tag{108}$$

$$-\frac{\alpha^2(\omega-1)\gamma_A\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)^2\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)+V^2\left(\gamma_A\tau_S+\tau_D\right)\right)}{\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)-1\right)-\pi_A V^2\left(\gamma_A\tau_S+\tau_D\right)\right)^2}. \tag{109}$$

A similar positive result holds with respect to the total volume of traders V:

$$\frac{\partial \text{Threshold}_\lambda}{\partial V} = \tag{110}$$

$$\frac{2\alpha^2\left(\left(-\pi_I\right)\right)\pi_A V(\omega-1)\gamma_A\sigma_u^2\left(\left(\tau_{\mu,A|s}\right)\right)\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)^2}{\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\sigma_\xi^2\left(\tau_A+\tau_D\right)-1\right)-\pi_A V^2\left(\left(\tau_{\mu,A|s}\right)\right)\right)^2}. \tag{111}$$

On the negative side, for risk aversion and noise trader standard deviation, the derivatives are negative and given by

$$\frac{\partial \text{Threshold}_{\lambda}}{\partial \alpha} = \tag{112}$$

$$- \frac{2\alpha\left(\left(-\pi_I\right)\right)\pi_A V^2(\omega-1)\gamma_A\left(\left(\tau_{\mu,A|s}\right)\right)\left(\sigma_\xi^2\sigma_u\left(\left(\tau_{\mu,A|s}\right)\right)+\sigma_u\right)^2}{\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\sigma_\xi^2\left(\tau_A+\tau_D\right)-1\right)-\pi_A V^2\left(\left(\tau_{\mu,A|s}\right)\right)\right)^2}. \tag{113}$$

$$\frac{\partial \text{Threshold}_{\lambda}}{\partial \sigma_u} = \tag{114}$$

$$- \frac{2\left(\left(-\pi_I\right)\right)\pi_A V^2(\omega-1)\gamma_A\sigma_u\left(\left(\tau_{\mu,A|s}\right)\right)\left(\alpha+\alpha\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)\right)^2}{\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\sigma_\xi^2\left(\tau_A+\tau_D\right)-1\right)-\pi_A V^2\left(\left(\tau_{\mu,A|s}\right)\right)\right)^2}. \tag{115}$$

.

C.2 Proposition 2.2

If we restrict to the space of positive profitability with $k > k^*$, the predictability is lower in the presence of more attentive agents, as shown by its derivative.

$$k > k^* \Rightarrow \frac{\partial \text{Profits}_{LLM}}{\partial \pi_A} = \tag{116}$$

$$(\alpha(\omega-1)\gamma_A\tau_D^2\tau_S\sigma_u^2\left(\sigma_\xi^2\left(\tau_A+\tau_D\right)+1\right)^2\times \tag{117}$$

$$\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)+V^2\left(\left(\tau_{\mu,A|s}\right)\right)\right)\times \tag{118}$$

$$(\lambda(c,k)\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\gamma_A\tau_S+\tau_D\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\tau_D^2\sigma_\xi^2+\tau_D\left(-\left(\left(\pi_A-1\right)(\omega-3)\gamma_A\sigma_\xi^2\tau_S\right)-1\right)+\tau_A\left(-\left(\left(\left(-\pi_I\right)\right)(\omega-2)\gamma_A\sigma_\xi^2\tau_S\right)-\pi_A\omega+\pi_A+\omega-2\right)\right)-\pi_A V^2\left(\gamma_A\tau_S+\tau_D\right)^2\right)+ \tag{119}$$

$$\gamma_A\left(\pi_A V^2\left(\left(\tau_{\mu,A|s}\right)\right)^2-\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(\left(\left(-\pi_I\right)\right)(2\omega-1)\tau_D^2\sigma_\xi^2+\tau_D\left(\left(\left(-\pi_I\right)\right)(3\omega-1)\gamma_A\sigma_\xi^2\tau_S+2\pi_A\omega-2\pi_A-2\omega+1\right)+\tau_A\left(\left(\left(-\pi_I\right)\right)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)-\pi_A\right)\right)\right))/ \tag{120}$$

$$(\left(\left(\tau_{\mu,A|s}\right)\right)\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\tau_D^2\sigma_\xi^2+\tau_D\left(\left(\pi_A-1\right)(\omega+1)\gamma_A\sigma_\xi^2\tau_S-1\right)+\tau_A\left(\left(\left(-\pi_I\right)\right)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)-\pi_A\right)\right)-\pi_A V^2\left(\tau_A+\tau_D\right)^2\right)^3)<0. \tag{121}$$

Conversely, any market where it is more costly to be an attentive agent will have a lower proportion of attentive agents.

C.3 Proposition 2.3

We have already shown above that an increase in the proportion of attentive agents, π_A, implies a decline in LLM profitability (if the profitability is positive) and in mispricing. In addition, if $k > k^*$, then higher volume, better information processing of inattentive agents, or lower noise risk reduces return predictability and mispricing, as shown by the following derivatives:

$$k > k^* \Rightarrow \frac{\partial \text{Threshold}_\lambda}{\partial V} = \tag{122}$$

$$(2\alpha\left((-\pi_I)\right)\pi_A V(\omega-1)\gamma_A\tau_D^2\tau_S\sigma_u^2\left(\sigma_\xi^2\left((\tau_{\mu,A|s})\right)+1\right)^2\times \tag{123}$$

$$(\lambda(c,k)\left(\pi_A V^2\left(\tau_A+\tau_D\right)^2-\alpha^2\sigma_u^2\left(\sigma_\xi^2\left((\tau_{\mu,A|s})\right)+1\right)\left((\pi_A-1)\tau_D^2\sigma_\xi^2+\tau_D\left(-\left(\left((-\pi_I)\right)(\omega-3)\gamma_A\sigma_\xi^2\tau_S\right)-1\right)+\tau_A\left(-\left(\left((-\pi_I)\right)(\omega-2)\gamma_A\sigma_\xi^2\tau_S\right)-\pi_A\omega+\pi_A+\omega-2\right)\right)+ \tag{124}$$

$$\gamma_A\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\gamma_A\tau_S+\tau_D\right)+1\right)\left(\left((-\pi_I)\right)(2\omega-1)\tau_D^2\sigma_\xi^2+\tau_D\left((\pi_A-1)(3\omega-1)\gamma_A\sigma_\xi^2\tau_S+2\pi_A\omega-2\pi_A-2\omega+1\right)+\tau_A\left(\left((-\pi_I)\right)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)-\pi_A\right)\right)-\pi_A V^2\left(\gamma_A\tau_S+\tau_D\right)^2\right)))/ \tag{125}$$

$$(\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\tau_A+\tau_D\right)+1\right)\left(\left((-\pi_I)\right)\tau_D^2\sigma_\xi^2+\tau_D\left(\left((-\pi_I)\right)(\omega+1)\gamma_A\sigma_\xi^2\tau_S-1\right)+\tau_A\left(\left((-\pi_I)\right)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)-\pi_A\right)\right)-\pi_A V^2\left((\tau_{\mu,A|s})\right)^2\right)^3) \tag{126}$$

$$< 0. \tag{127}$$

$$k > k^* \Rightarrow \frac{\partial \text{Threshold}_\lambda}{\partial \sigma_u} = \tag{128}$$

$$(2\alpha\left((-\pi_I)\right)\pi_A V^2(\omega-1)\tau_A\sigma_u\left(\tau_D\sigma_\xi^2\left(\gamma_A\tau_S+\tau_D\right)+\tau_D\right)^2\times \tag{129}$$

$$(\lambda(c,k)(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\gamma_A\tau_S+\tau_D\right)+1\right)\times \tag{130}$$

$$\left(\left((-\pi_I)\right)\tau_D^2\sigma_\xi^2+\tau_D\left(-\left((\pi_A-1)(\omega-3)\gamma_A\sigma_\xi^2\tau_S\right)-1\right)+\tau_A\left(-\left(\left((-\pi_I)\right)(\omega-2)\gamma_A\sigma_\xi^2\tau_S\right)-\pi_A\omega+\pi_A+\omega-2\right)\right)- \tag{131}$$

$$\pi_A V^2\left(\gamma_A\tau_S+\tau_D\right)^2)+\gamma_A\times \tag{132}$$

$$\left(\pi_A V^2\left((\tau_{\mu,A|s})\right)^2-\alpha^2\sigma_u^2\left(\sigma_\xi^2\left((\tau_{\mu,A|s})\right)+1\right)\left(\left((-\pi_I)\right)(2\omega-1)\tau_D^2\sigma_\xi^2+\tau_D\left(\left((-\pi_I)\right)(3\omega-1)\gamma_A\sigma_\xi^2\tau_S+2\pi_A\omega-2\pi_A-2\omega+1\right)+\tau_A\left(\left((-\pi_I)\right)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)-\pi_A\right)\right)\right)/ \tag{133}$$

$$(\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left((\tau_{\mu,A|s})\right)+1\right)\left(\left((-\pi_I)\right)\tau_D^2\sigma_\xi^2+\tau_D\left(\left((-\pi_I)\right)(\omega+1)\gamma_A\sigma_\xi^2\tau_S-1\right)+\tau_A\left(\left((-\pi_I)\right)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)-\pi_A\right)\right)-\pi_A V^2\left((\tau_{\mu,A|s})\right)^2\right)^3) \tag{134}$$

$$> 0. \tag{135}$$

$$k > k^* \Rightarrow \frac{\partial \text{Threshold}_\lambda}{\partial \omega} = \tag{136}$$

$$(\alpha\left(\left(-\pi_I\right)\right)\gamma_A\tau_D^2\tau_S\sigma_u^2\left(\sigma_\xi^2\left(\tau_A+\tau_D\right)+1\right)^2\times \tag{137}$$

$$(\lambda(c,k)\left(\pi_A V^2\left(\left(\tau_{\mu,A|s}\right)\right)-\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\sigma_\xi^2\left(\gamma_A\tau_S+\tau_D\right)-1\right)\right)\times \tag{138}$$

$$\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\tau_A+\tau_D\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\tau_D^2\sigma_\xi^2+\tau_D\left(-\left(\left(\left(-\pi_I\right)\right)(\omega-3)\gamma_A\sigma_\xi^2\tau_S\right)-1\right)+\tau_A\left(-\left(\left(\left(-\pi_I\right)\right)(\omega-2)\gamma_A\sigma_\xi^2\tau_S\right)-\pi_A\omega+\pi_A+\omega-2\right)\right)-\pi_A V^2\left(\tau_A+\tau_D\right)^2\right)+ \tag{139}$$

$$\gamma_A\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)-1\right)-\pi_A V^2\left(\gamma_A\tau_S+\tau_D\right)\right)\times \tag{140}$$

$$\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(\left(\left(-\pi_I\right)\right)(2\omega-1)\tau_D^2\sigma_\xi^2+\tau_D\left(\left(\left(-\pi_I\right)\right)(3\omega-1)\gamma_A\sigma_\xi^2\tau_S+2\pi_A\omega-2\pi_A-2\omega+1\right)+\tau_A\left(\left(\left(-\pi_I\right)\right)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)-\pi_A\right)\right)-\pi_A V^2\left(\tau_A+\tau_D\right)^2\right)))/ \tag{141}$$

$$(\left(\left(\tau_{\mu,A|s}\right)\right)\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\tau_D^2\sigma_\xi^2+\tau_D\left((\pi_A-1)(\omega+1)\gamma_A\sigma_\xi^2\tau_S-1\right)+\tau_A\left(\left(\left(-\pi_I\right)\right)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)-\pi_A\right)\right)-\pi_A V^2\left(\tau_A+\tau_D\right)^2\right)^3) \tag{142}$$

$$<0. \tag{143}$$

D Theorem 2

In the case of inattentive agents using LLMs, the price is given by:

$$p_{1|I}=\frac{-\left(\left(-\pi_I\right)\right)t\left(\bar{d}\tau_D+s\tau_S\lambda(c,k)\right)+\left(\left(-\pi_I\right)\right)(t-1)\left(\bar{d}\tau_D+s\omega\tau_A\right)+\frac{\pi_A(\bar{d}\tau_D+s\tau_A)\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)+V^2\left(\left(\tau_{\mu,A|s}\right)\right)\right)}{\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)^2}}{-\left(\left(-\pi_I\right)\right)t\left(\tau_S\lambda(c,k)+\tau_D\right)+\left(\left(-\pi_I\right)\right)(t-1)\left(\omega(\tau_{\mu,A|s})\right)+\frac{\pi_A\left(\frac{1}{(\tau_{\mu,A|s})}+\sigma_\xi^2+\frac{V^2}{\alpha^2\sigma_u^2}\right)}{\left(\frac{1}{\gamma_A\tau_S+\tau_D}+\sigma_\xi^2\right)^2}}. \tag{144}$$

Hence, mispricing in this case is:

$$E[\alpha_{M|I}^2]= \tag{145}$$

$$(\tau_D^2\tau_S(-\alpha^2\left(\left(-\pi_I\right)\right)t\sigma_u^2\lambda(c,k)\left(\sigma_\xi^2\left(\gamma_A\tau_S+\tau_D\right)+1\right)^2+ \tag{146}$$

$$\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\tau_D\sigma_\xi^2+\gamma_A\left(\left(\left(-\pi_I\right)\right)\left((t-1)\omega\left(\tau_D\sigma_\xi^2+1\right)+\sigma_\xi^2\tau_S\right)+\pi_A\right)+\left(\left(-\pi_I\right)\right)(t-1)\omega\gamma_A^2\sigma_\xi^2\tau_S-1\right) \tag{147}$$

$$+\pi_A V^2\left(\gamma_A-1\right)\left(\left(\tau_{\mu,A|s}\right)\right))^2)/ \tag{148}$$

$$(\left(\tau_D+\tau_S\right)^2(\alpha^2\left(\left(-\pi_I\right)\right)t\tau_S\sigma_u^2\lambda(c,k)\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)^2+ \tag{149}$$

$$\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\tau_A+\tau_D\right)+1\right)\left(\left(\left(-\pi_I\right)\right)\tau_D^2\sigma_\xi^2+\tau_D\left(-\left(\left(\left(-\pi_I\right)\right)\gamma_A\sigma_\xi^2\tau_S((t-1)\omega-1)\right)-1\right)-\tau_A\left(\left(\left(-\pi_I\right)\right)(t-1)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)+\pi_A\right)\right)-\pi_A V^2\left(\tau_A+\tau_D\right)^2)^2). \tag{150}$$

The derivative of mispricing w.r.t. the proportion of inattentive agents using LLMs, θ, is negative:

$$\frac{\partial E[\alpha^2_{M|I}]}{\partial \theta} = \tag{151}$$

$$(2\alpha^2\left((-\pi_I)\right)\tau_D^2\tau_S\sigma_u^2\left(\omega\gamma_A-\lambda(c,k)\right)\times \tag{152}$$

$$\left(\sigma_\xi^2\left((\tau_{\mu,A|s})\right)+1\right)^2\times\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\gamma_A\tau_S+\tau_D\right)+1\right)\left(\left((-\pi_I)\right)\sigma_\xi^2\left((\tau_{\mu,A|s})\right)-1\right)-\pi_A V^2\left((\tau_{\mu,A|s})\right)\right)\times \tag{153}$$

$$(-\alpha^2\left((-\pi_I)\right)t\sigma_u^2\lambda(c,k)\left(\sigma_\xi^2\left((\tau_{\mu,A|s})\right)+1\right)^2+ \tag{154}$$

$$\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\gamma_A\tau_S+\tau_D\right)+1\right)\left(\left((-\pi_I)\right)\tau_D\sigma_\xi^2+\gamma_A\left(\left((-\pi_I)\right)\left((t-1)\omega\left(\tau_D\sigma_\xi^2+1\right)+\sigma_\xi^2\tau_S\right)+\pi_A\right)+\left((-\pi_I)\right)(t-1)\omega\gamma_A^2\sigma_\xi^2\tau_S-1\right)+\pi_A V^2\left(\gamma_A-1\right)\left(\tau_A+\tau_D\right)))/ \tag{155}$$

$$\left(\tau_D+\tau_S\right)(\alpha^2\left((-\pi_I)\right)t\tau_S\sigma_u^2\lambda(c,k)\left(\sigma_\xi^2\left((\tau_{\mu,A|s})\right)+1\right)^2+ \tag{156}$$

$$\alpha^2\sigma_u^2\left(\sigma_\xi^2\left((\tau_{\mu,A|s})\right)+1\right)\left(\left((-\pi_I)\right)\tau_D^2\sigma_\xi^2+\tau_D\left(-\left(\left((-\pi_I)\right)\gamma_A\sigma_\xi^2\tau_S((t-1)\omega-1)\right)-1\right)-\tau_A\left(\left((-\pi_I)\right)(t-1)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)+\pi_A\right)\right)-\pi_A V^2\left((\tau_{\mu,A|s})\right)^2)^3) \tag{157}$$

$$<0. \tag{158}$$

The derivative of mispricing w.r.t. the LLM's model size, k, is also negative:

$$\frac{\partial E[\alpha^2_{M|I}]}{\partial k} = \tag{159}$$

$$-(2\alpha^2\left((-\pi_I)\right)t\tau_D^2\tau_S\sigma_u^2\frac{\partial\lambda}{\partial k}\left(\sigma_\xi^2\left(\gamma_A\tau_S+\tau_D\right)+1\right)^2\times \tag{160}$$

$$\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\tau_A+\tau_D\right)+1\right)\left(\left((-\pi_I)\right)\sigma_\xi^2\left((\tau_{\mu,A|s})\right)-1\right)-\pi_A V^2\left((\tau_{\mu,A|s})\right)\right)\times \tag{161}$$

$$(-\alpha^2\left((-\pi_I)\right)t\sigma_u^2\lambda(c,k)\left(\sigma_\xi^2\left((\tau_{\mu,A|s})\right)+1\right)^2+ \tag{162}$$

$$\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\gamma_A\tau_S+\tau_D\right)+1\right)\left(\left((-\pi_I)\right)\tau_D\sigma_\xi^2+\gamma_A\left(\left((-\pi_I)\right)\left((t-1)\omega\left(\tau_D\sigma_\xi^2+1\right)+\sigma_\xi^2\tau_S\right)+\pi_A\right)+\left((-\pi_I)\right)(t-1)\omega\gamma_A^2\sigma_\xi^2\tau_S-1\right)+\pi_A V^2\left(\gamma_A-1\right)\left(\tau_A+\tau_D\right)))/ \tag{163}$$

$$\left((\tau_D+\tau_S\right)(\alpha^2\left((-\pi_I)\right)t\tau_S\sigma_u^2\lambda(c,k)\left(\sigma_\xi^2\left((\tau_{\mu,A|s})\right)+1\right)^2+ \tag{164}$$

$$\alpha^2\sigma_u^2\left(\sigma_\xi^2\left((\tau_{\mu,A|s})\right)+1\right)\left(\left((-\pi_I)\right)\tau_D^2\sigma_\xi^2+\tau_D\left(-\left(\left((-\pi_I)\right)\gamma_A\sigma_\xi^2\tau_S((t-1)\omega-1)\right)-1\right)-\tau_A\left(\left((-\pi_I)\right)(t-1)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)+\pi_A\right)\right)-\pi_A V^2\left((\tau_{\mu,A|s})\right)^2)^3) \tag{165}$$

$$<0. \tag{166}$$

D.1 Proposition 2.4

The profitability of LLM strategies when inattentive agents is given by

$$\text{Profitability}_{LLM|I} \equiv \tag{167}$$

$$(\alpha\left(\left(-\pi_I\right)\right)\tau_D^2\tau_S\sigma_u^2\left(\sigma_\xi^2\left(\tau_A+\tau_D\right)+1\right)^2\times \tag{168}$$

$$\left(-t\lambda(c,k)+\gamma_A+(t-1)\omega\gamma_A\right)\left(\gamma_A\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(\left(\left(-\pi_I\right)\right)(t-1)\omega\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)+\pi_A\right)+\pi_A V^2\left(\tau_A+\tau_D\right)\right)- \tag{169}$$

$$\lambda(c,k)\left(\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\left(-\left(\left(\left(-\pi_I\right)\right)\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)\right)+\left(\left(-\pi_I\right)\right)t\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)+1\right)+\pi_A V^2\left(\tau_A+\tau_D\right)\right)\Big)/ \tag{170}$$

$$\left(\left(\left(\tau_{\mu,A|s}\right)\right)\left(\alpha^2\left(\left(-\pi_I\right)\right)t\tau_S\sigma_u^2\lambda(c,k)\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)^2+\right. \tag{171}$$

$$\alpha^2\sigma_u^2\left(\sigma_\xi^2\left(\left(\tau_{\mu,A|s}\right)\right)+1\right)\times \tag{172}$$

$$\left.\left(\left(\left(-\pi_I\right)\right)\tau_D^2\sigma_\xi^2+\tau_D\left(-\left(\left(\left(-\pi_I\right)\right)\gamma_A\sigma_\xi^2\tau_S\left((t-1)\omega-1\right)\right)-1\right)-\tau_A\left(\left(\left(-\pi_I\right)\right)(t-1)\omega\left(\gamma_A\sigma_\xi^2\tau_S+1\right)+\pi_A\right)-\pi_A V^2\left(\left(\tau_{\mu,A|s}\right)\right)^2\right)^2\Big). \tag{173}$$

And the proposition follows immediately by substituting $\lambda(c,k')=\gamma$ for large enough k', and $\theta=1$.

E Proposition 2.5

This proposition follows immediately from Proposition 2.1 because the parameter γ is substituted by $\lambda(c,k)$.

Appendix E: Topic Modeling

In this section, we discuss more details about our topic modeling technique. Note that we focus only on non-neutral GPT scores, which leaves 87,699 data points in this analysis. We also remove the 'YES,' 'NO', or 'UNKNOWN' from the explanations to avoid any mechanical effect.

We fit the BERTopic model with standard parameters (10 top words). We use Sentence-Transformer from Hugging Face for the embeddings with the model "all-MiniLM-L6-v2" and use UMAP (McInnes et al. (2018)) for dimensionality reduction with 5 dimensions, 15 neighbors, and a minimum distance of 0. In addition, we use HDBSCAN (McInnes, Healy, and Astels (2017)) for clustering with a minimum cluster size of 100, which in practice means it will not consider topics that are prevalent in less than 100 data points. We do not include firm or date fixed effects to ease the interpretation, but standard errors are clustered by firm and date. The exercise results in 60 topics.

Lastly, we use the labels generated by ChatGPT 4 based on the topic model using the following prompt:

> You are tasked with extracting a short but highly descriptive topic label from a topic model and a set of example documents from that topic.
>
> The documents contain [DOCUMENT TYPE].
>
> Follow these instructions carefully:
>
> 1. Review the following documents: [DOCUMENTS]
>
> 2. Consider the following keywords that describe the topic: [KEYWORDS]
>
> 3. Based on the information provided in the documents and keywords, extract a short but highly descriptive topic label.
>
> a. Keep the label at most 4 words long.
>
> b. Do NOT include specific names, firm names, or stock tickers in the label, as the examples may come from only a few companies.
>
> c. Interpret the concise topic rather than focusing on specific entities.
>
> d. Make the label MORE specific than general terms like 'Financial Results', 'Earnings Announcement', 'Financial Announcements', 'Financial Updates', or 'Corporate Finance Announcements.'
>
> e. Avoid using firm names in whole or in part, tickers, and the words 'Quarterly' or 'Financial' in your label.

f. Focus on the specific theme that best describes the collection of explanations and keywords.

Keep your topic label concise, specific, and descriptive.

Provide your final topic label in the following format: topic: <topic label>