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

Alejandro Lopez-Lira and Yuehua Tang

University of Florida

First Version: April 6, 2023; This Version April 17, 2024

Abstract

We examine the potential of ChatGPT and other large language models (LLMs) to predict stock market returns using news headlines. ChatGPT scores significantly predict subsequent daily stock returns, outperforming traditional methods. A model involving information capacity constraints, limits to arbitrage, and LLMs rationalizes this predictability, which strengthens among smaller stocks and following negative news. The model predicts a distinct impact of LLMs on market efficiency relative to other machine-learning developments due to their simultaneous wide availability. Only advanced LLMs maintain accuracy when interpreting hard-to-read news, while basic models cannot accurately forecast returns, suggesting return forecasting is an emerging capacity of bigger LLMs, which deliver higher Sharpe ratios. We develop an interpretability technique to evaluate LLMs' reasoning.

Keywords: Large Language Models, ChatGPT, Machine Learning, Return Predictability, Textual Analysis
JEL Classification: C81, G11, G12, G18

The recent proliferation of generative artificial intelligence and large language models (LLMs) like ChatGPT has sparked interest in their potential applications. However, their capabilities in financial economics remain relatively unexplored, particularly regarding predicting price movements. We document that the latest LLMs demonstrate out-of-sample proficiency for return prediction when leveraging news, subsuming traditional methods, and demonstrating the potential value of LLMs in financial tasks for asset managers interested in using them. For instance, without considering transaction costs, a self-financing strategy that buys stocks with a positive ChatGPT recommendation and sells stocks with negative recommendations earns a cumulative return of over 400% from October 2021 to December 2022.[1] Rather than optimize this strategy for implementation, we focus on the economic insights derived from the analysis.

Extensive literature documents that stock returns are predictable at a daily horizon using supervised textual analysis algorithms.[2] Unlike prior literature, our study takes a novel approach by evaluating whether LLMs can acquire return prediction abilities *despite* no explicit financial training. LLMs are not trained to forecast returns, but their training on massive text data enables abilities from nuanced language comprehension to emergent reasoning capacities. Besides their lack of explicit training, LLMs' financial capabilities are intriguing due to their general availability, as their accessibility enables a wide range of individuals and organizations to leverage them, potentially reducing the computational advantages of large institutions and facilitating greater market efficiency.

LLMs exhibiting proficiency in return prediction even in recent years presents a puzzle. To solve this puzzling predictability, we develop a theoretical model with information processing constraints, limits to arbitrage, and LLMs. To the best of our knowledge, this is one of the first models to incorporate LLMs in economic settings. This theoretical framework generates several testable implications that we confirm empirically.

First, the return predictability aligns with theories of delayed information diffusion,

---

1. Assuming a transaction cost of 10 (25) basis points per trade, the strategy earns a cumulative return of 350% (50%) over our sample period. 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 various risk factors.

2. See, for example, Tetlock (2007), Tetlock, Saar-Tsechansky, and Macskassy (2008), and Tetlock (2011), among others.

bounded investor attention, and limits to arbitrage. Second, greater model complexity enhances information processing, so more advanced LLMs better forecast returns. In fact, only models passing a certain capacity threshold can predict returns with the correct sign. Moreover, only sophisticated LLMs can exploit low-readability news. Third, our model predicts increased price informativeness once LLMs are sophisticated enough and agents start using them. However, a persistent pattern of return predictability remains in equilibrium. Its magnitude depends on the non-fundamental volatility and transaction fees. Moreover, we study the implications of LLMs exhibiting better-than-human information processing capabilities.

Empirically, basic models like GPT-1, GPT-2, and BERT display little stock forecasting capabilities and no significant return predictability. In contrast, we observe the highest predictability in the more complex models like ChatGPT-4. A self-financing strategy that buys the stocks with a positive ChatGPT-4 score and sells stocks with a negative ChatGPT-4 score after the news announcement delivers the highest Sharpe ratio of 3.8 over our sample period, compared to a Sharpe ratio of 3.1 for the strategy based on ChatGPT-3.5.

Furthermore, the predictability of the ChatGPT scores is present among both small and large-cap stocks and in stocks with positive and negative news. Our results suggest that the market may be underreacting to company daily news, consistent with the evidence in the extant literature.[3] Nevertheless, the predictability is more pronounced among smaller stocks. For instance, the coefficient on the ChatGPT score in predicting returns of small stocks (i.e., less than the 10th percentile NYSE market cap distribution) is more than four times the magnitude of the one for the remaining sample. In addition, stocks with negative news headlines also show stronger predictability.

Next, we examine LLMs' forecasting capabilities across headlines with different complexity levels. We categorize news into low- or high-difficulty using the Flesch-Kincaid Readability Score and study how diverse LLMs perform when news is more difficult to understand. Advanced LLMs like GPT-4 can predict returns in low-readability news, unlike more basic models that fail at this task. Extracting helpful trading signals from low-readability news highlights the unique advantage of advanced LLMs in financial markets.

---

3. For example, Bernard and Thomas (1989), Chan, Jegadeesh, and Lakonishok (1996), and Jiang, Li, and Wang (2021), Fedyk and Hodson (2023))

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 and FinBERT, 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 GPT-4 exhibit sustained predictive accuracy, indicating their resilience to potential biases in these communications. These biases are especially pertinent for press releases, which may be crafted to influence stakeholder perceptions. This advancement suggests that the risk of market manipulation via corporate communication may become less critical to regulators as these models can identify subtleties and biases in firm-issued statements (Cao et al. (2023)).

Finally, we propose a new method to evaluate and understand the models' reasoning capabilities using these models' ability to explain their reasoning. First, we assess whether the recommendations (excluding the neutral ones) were correct by comparing them with the realized return the following day. Second, we fit a model that predicts whether a recommendation is correct or incorrect using the recommendation explanation. Third, we use the model's feature importance to understand which concepts are better at predicting correct recommendations.[4] We look at the most important words and their surrounding context and find that the answers are more likely to be correct when the reasoning is related to stock purchases by insiders, earning guidance, and dividends. On the contrary, the model does worse when its logic refers to partnerships and developments.

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 abstain from studying in this paper. Second, the success of LLMs in this complex, noisy domain underscores the emergent advanced reasoning capabilities of state-of-the-art AI systems in economic tasks. Third, our study can benefit asset managers and institutional investors by providing empirical evidence on the efficacy of LLMs in predicting stock market returns. Fourth, regulators have to be aware LLMs' influence on market behavior, information dissemination, and price formation

---

4. Feature importance refers to how relevant a specific variable is for a model to predict correctly. When a variable with high feature importance is removed, the model will predict worse. See Binsbergen, Han, and Lopez-Lira (2023) for a detailed exposition.

will become critical areas of concern. Finally, understanding LLMs' ability to forecast returns sheds light on market efficiency, information processing, and limits to arbitrage.

Related Literature

Recent papers that use ChatGPT in the context of economics include Hansen and Kazinnik (2023), Cowen and Tabarrok (2023), Korinek (2023), and Noy and Zhang (2023). Hansen and Kazinnik (2023) show that LLMs like ChatGPT can decode Fedspeak (i.e., the language used by the Fed to communicate on monetary policy decisions). Cowen and Tabarrok (2023) and Korinek (2023) demonstrate that ChatGPT is helpful in teaching economics and conducting economic research. Noy and Zhang (2023) find that ChatGPT can enhance productivity in professional writing jobs. Furthermore, Yang and Menczer (2023) demonstrates that Chat-GPT successfully identifies credible news outlets. Contemporaneously, 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. We attribute the difference in results to their focus on using historical numerical data to predict, while ChatGPT excels at textual tasks. Our study is among the first to document the potential of LLMs in financial markets.

We also contribute to the recent strand of the literature that employs textual analysis and machine learning to study a variety of finance research questions (e.g., Jegadeesh and Wu (2013), Rapach, Strauss, and Zhou (2013), Campbell et al. (2014), Hoberg and Phillips (2016), Gaulin (2017), Baker, Bloom, and Davis (2016), Manela and Moreira (2017), Hansen, McMahon, and Prat (2018), Ke, Kelly, and Xiu (2019), Ke, Montiel Olea, and Nesbit (2019), Bybee et al. (2019), F. Jiang et al. (2019), Gu, Kelly, and Xiu (2020), Cohen, Malloy, and Nguyen (2020), Freyberger, Neuhierl, and Weber (2020), Lopez-Lira (2019), Binsbergen et al. (2020), Bybee et al. (2021), Chin and Fan (2023)). Our paper also adds the literature that uses linguistic analyses of news articles to extract sentiment and predict stock returns. One strand of this literature studies media sentiment and aggregate stock returns (e.g., Tetlock (2007), Garcia (2013), Calomiris and Mamaysky (2019)). Another strand uses the sentiment of firm news to predict future individual stock returns (e.g., Tetlock, Saar-Tschansky, and Macskassy (2008), Tetlock (2011), Jiang, Li, and Wang (2021)). Different from prior stud-

ies, we focus on understanding whether LLMs add value by extracting signals from news headlines that predict stock market reactions. Our unique contribution is being the first to provide a theoretical framework and comprehensive empirical evidence on large language models' stock return predictability skills.

Our paper also relates to the literature on employment exposures and vulnerability to AI-related technology by documenting a task where LLMs perform well. Recent works by Agrawal, Gans, and Goldfarb (2019), Webb (2019), Acemoglu et al. (2022), Acemoglu and Restrepo (2022), Babina et al. (2022), W. Jiang et al. (2022), and Noy and Zhang (2023) have examined the extent of job exposure and vulnerability to AI-related technology as well as the consequences for employment and productivity.

Finally, our research contributes to the broader academic discourse on artificial intelligence applications in finance. Understanding LLMs' potential when predicting stock market returns can inspire future research on developing LLMs tailored to the financial industry's needs.[5]

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 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. Using unsupervised learning methods, it is pre-trained on a large corpus of text data, such as Wikipedia articles or web pages. This pre-training

---

5. See for example Wu et al. (2023) on "BloombergGPT: A Large Language Model for Finance."

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-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. By considering the more basic models alongside Chat-GPT, we can examine the importance of model complexity for stock return prediction based on textual data. Online Appendix A gives an overview of the nine LLMs we study in this paper, ordered by their release date.

2   Model

We use an economic model to understand the sources of return predictability and why more sophisticated language models can better exploit more complex information contained in the news. We work in a similar setup as Gromb and Vayanos (2010) model, but we add an information structure to mimic unexpected news releases that provide information about the asset's fundamentals, LLMs, and slow-moving expectations for a subset of the agents due to attention constraints.

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 we can focus on expectations rather than risk premiums.

There is a total population of agents with measure $V > 0$. V helps us model the

total trading of the asset, and we can think of it as the volume from active traders without considering the nonfundamental demand. Hence, we would expect V to be lower in smaller stocks and, by definition, in illiquid stocks. Attentive agents have a total measure of $\pi_A V, \pi_A \in (0, 1)$ and inattentive agents a measure of $\pi_I V, \pi_I = 1 - \pi_A$. π_A is determined in equilibrium and depends on the costs of participating in the market as an attentive investor and the attractiveness of speculation. We assume these costs depend implicitly on the firm size and whether attentive agents need to short the stock or not. Hence, having two types of agents allows us to explore the impact of different firm characteristics in a reduced form specification.

The model features three periods. In period one, unexpected information about the company is revealed. Attentive agents incorporate this information in their forecasts better than inattentive investors. We assume inattentive agents only update 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 the dividend, $\tilde{d}$, and the economy ends. There is no intertemporal discount for notational ease because 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. Additional sources of risk can include the possibility of further news within the same period, inattentive agents not fully updating their beliefs with some probability, or nonfundamental risk stemming from noise traders. We settle for the latter due to tractability and because it closely maps the market environment. When trading within a day, there is a chance nonfundamental trades will move the price further away from its fundamental value.

Asset with uncertain dividends   Information revealed

	No news		Partial update by inattentive		Full update	
Period 0	$\longrightarrow$	Period 1	$\longrightarrow$	Period 2	$\longrightarrow$	Period 3

Two types of agents   Inattentive partially updates   Inattentive fully update

Attentive fully updates

2.1 Equilibrium Without News

The asset payoff is random and, without news, 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 is uncertainty because agents do not know the fundamentals, and there is real variation in the asset's payoff given by $\sigma_\xi \tilde{\xi}$.

Agents do not know the fundamentals but have a correct prior. The prior is correct in the sense that

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

We assume the uncertainty about the fundamentals and the real variation are independent.

We let $x_{j,t+1}$ denote the gross demand of the asset for each type of investor in period t. In period zero, agents maximize

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

Without unexpected news, the demand of both agents is given by:

$$x_{1,j} = \frac{\bar{d} - p_0}{\alpha(\sigma_d^2 + \sigma_\xi^2)}. \tag{4}$$

The asset is in zero net supply. Given noise trader demand (u), we obtain the equilibrium

price by the market clearing condition:

$$V \frac{\bar{d} - p_0}{\alpha(\sigma_d^2 + \sigma_\xi^2)} + u = 0 \tag{5}$$

The price at period zero and its (unconditional) expected value are given by:

$$p_0 = \bar{d} + u \frac{\alpha(\sigma_d^2 + \sigma_\xi^2)}{V}, \ \ E[p_0] = \bar{d} \tag{6}$$

2.2 Information Structure

Unexpectedly, news about the company gets released. News can either be interpreted as a change in a firm's fundamentals or as new information about a firm's fundamentals.

Agents interpret this news as signals with information about the fundamentals. The amount of information they process is related to their information processing capacity, the complexity of the news, and the LLMs' size.

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

$$s = \mu_d + \varepsilon, \quad \varepsilon \sim N(0, \sigma_S^2 = \frac{1}{\tau_S}) \tag{7}$$

All agents, including attentive, inattentive, and large language models (LLMs), observe the same signal realization s, representing the company's news. 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 that everyone observes the same signal. We could incorporate bias at the cost of additional notation.[6]

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

$$\tau_A = \gamma \tau_S, \ \ \tau_I = \omega \tau_A, \ \ \tau_L = \lambda(c, k)\tau_S \tag{8}$$

---

6. 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 potentially different return dynamics at the cost of a more complex model structure.

where $\gamma \in (0,1]$ is attentive agents' information capacity and $\omega \in (0,1]$ are exogenous constants. $\lambda(c,k) : \mathbb{R}_2^+ \to (0,1]$ is a twice differentiable, decreasing function of news complexity c and an increasing function of model size k, and represents LLMs' information . 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 capacity of LLMs decreases. However, as the model size of LLMs increases, their information processing capacity improves, even in the presence of complex news.

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 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 full implications of the news and have the chance to trade on them again. The signal inattentive investors processed is entirely contained in the signal the attentive investors processed. For example, some information may be complex to understand, and we assume attentive investors better understand this information.

Attentive agents only process up to 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 point of view of an agent outside the economy who can process the entire signal. We will study this case later by incorporating a hypothetical LLM with a large 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, and intuitively we are thinking that a human would revise the output. Otherwise, we assume there is just one update.

We further assume that inattentive investors do not condition on prices. Attentive investors have the complete information set available in the economy, so they do not need to learn from the price if LLMs are worse than them at information processing. Only when we introduce advanced LLMs with information capacity greater than attentive agents will they

11

learn from the prices.

2.3 Information Update

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

$$\mu_{j|s} \equiv E[d|s_j] = \frac{\bar{d}\tau_d + s_j\tau_{1,j}}{\tau_d + \tau_{1,j}}, j \in \{A, I, LLM\}. \tag{9}$$

The updated fundamental variance is given by

$$\sigma_{j|s} \equiv \frac{1}{\tau_{j|s}} \equiv Var[d|s_j] = \frac{1}{\tau_d + \tau_{1,j}}, \ j \in \{A, I, LLM\} \tag{10}$$

2.4 No Use of LLMs

We first focus on the case where agents are not using LLMs to trade. This case reflects the period when LLMs were unavailable. Intuitively, we are modeling the case of how LLMs would have reacted (out-of-sample) to the information. 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.

2.5 Period Two

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:

$$\mu_{2,j} = \mu_A \equiv E[\tilde{d}|s_A] = \frac{\bar{d}\tau_d + s\tau_A}{\tau_d + \tau_A}, \ j \in \{A, I\}. \tag{11}$$

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

$$\sigma_{2,j}^2 = \sigma_{A|s}^2 \equiv \frac{1}{\tau_{A|s}} \equiv Var[d|s_A] = \frac{1}{\tau_d + \tau_A}, \ j \in \{A, I\}. \tag{12}$$

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:

$$x_{3,j} = x_3 \equiv \frac{\mu_A - p_2}{\alpha(\sigma^2_{A|s} + \sigma^2_\xi)} \quad j \in \{A, I\}. \tag{13}$$

There are nonfundamental traders in the second period. We assume the demand from noise traders (u) is normally distributed with an expected value of zero and variance of σ^2_u

The market clearing condition is given by:

$$V\pi_A x_3 + V\pi_I x_3 + u = V\pi_A x_3 + V(1 - \pi_A)x_3 + u_2 = Vx_3 + u_2 = 0. \tag{14}$$

Hence, the price is given by

$$p_2 = \mu_A + u_2 \frac{\alpha(\sigma^2_{A|s} + \sigma^2_\xi))}{V}. \tag{15}$$

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

$$E_{1,A}[p_2] = E_{1,A}[\mu_A + u_2 \frac{\alpha(\sigma^2_{A|s} + \sigma^2_\xi)}{V}] = \mu_A = E[\tilde{d}|s_A]. \tag{16}$$

The expectation of time two is measurable at time 1 since there are no further updates on attentive agents. Hence, the only source of variance in the price is the nonfundamental demand, which is normally distributed with variance given by

$$\sigma^2_p \equiv Var_{1,A}(p_2) = \alpha^2 \frac{\sigma^2_u}{V^2}(\sigma^2_{A|s} + \sigma^2_\xi)^2. \tag{17}$$

$$\tau_p \equiv \frac{1}{\sigma^2_p} = \frac{1}{Var_{1,A}(p_2)} = \frac{V\tau_u * \tau^2_{A|s}}{\alpha^2}. \tag{18}$$

The expected price is equal to the expected dividend, while 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   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.

2.6.1   Attentive agents

Let $\tilde{w}_3 = w_2 + x(\tilde{d} - p_2)$. Given the demand in period 2 and the equilibrium prices, the value function of period 2 for attentive agents is:

$$V2(w_2) = \max_{x} E[U(\tilde{w}_3)] = \max_{x_3} -exp\{-\alpha E[\tilde{w}_3] + \frac{\alpha^2}{2} Var(\tilde{w}_3)\} \tag{19}$$

$$= -exp\{-\alpha(w_2) - \frac{1}{2} \frac{(\mu_A - p_2)^2}{(\sigma_{A|s}^2 + \sigma_\xi^2)}\} \tag{20}$$

Let $\tilde{w}_2 = w_1 + x_2(\tilde{p}_2 - p_1)$. In period 1, attentive agents maximize the expected period-2 value function:

$$V1(w_1) = \max_{x_2} E_1[V2(\tilde{w}_2)] = \max_{x} -E_1[exp\{-\alpha(w_2) - \frac{1}{2} \frac{(\mu_A - \tilde{p}_2)^2}{(\sigma_{A|s}^2 + \sigma_\xi^2)}\}] \tag{21}$$

$$= \max_{x} -E_1[exp\{a(x) + b(x)\tilde{p}_2 + c\tilde{p}_2^2\}] \tag{22}$$

with $a = a(x) = \alpha p_1 x - \frac{\mu_A^2}{2(\sigma_{A|s}^2 + \sigma_\xi^2)} - \alpha w_1$, $b = b(x) = \frac{\mu_A}{(\sigma_{A|s}^2 + \sigma_\xi^2)} - \alpha x$, and $c = c(x) = -\frac{1}{2(\sigma_{A|s}^2 + \sigma_\xi^2)}$.

Since $p_2 \sim N(\mu_p \equiv \mu_A, \sigma_p^2 \equiv \alpha^2 \frac{\sigma_u^2}{V^2}((\sigma_{A|s}^2 + \sigma_\xi^2)^2))$, we use the following lemma to obtain a closed form solution:

Lemma 2.1. *Let $z \sim N(\mu, \sigma^2)$. Then, if $1 - 2c\sigma^2 > 0$,*

$$E[exp(a + bz + cz^2)] = \frac{exp\left(-\frac{\sigma^2\left(b^2 - 4ac\right) + 2(a + \mu(b + c\mu))}{4c\sigma^2 - 2}\right)}{\sqrt{1 - 2c\sigma^2}}. \tag{23}$$

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

14

It follows that

$$V1(w_1) = \max_x \frac{\exp\left(\frac{-4ac\sigma_p^2 + 2a + b^2\sigma_p^2 + 2b\mu_p + 2c\mu_p^2}{2 - 4c\sigma_p^2}\right)}{\sqrt{1 - 2c\sigma_p^2}},$$ (24)

with the appropriate substitution of a, b, and c.

Hence, we can then obtain the optimal demand from the first-order condition.

$$x_{2,A} = \frac{(\mu_A - p_1)\left(\alpha^2(\sigma_{A|s}^2 + \sigma_\xi^2)\frac{\sigma_u^2}{V^2} + 1\right)}{\alpha^3(\sigma_{A|s}^2 + \sigma_\xi^2)^2\frac{\sigma_u^2}{V^2}}.$$ (25)

Notice that we can write the demand function as

$$x_{2,A} = \frac{(\mu_A - p_1)}{\alpha(\sigma_{A|s}^2 + \sigma_\xi^2)} + \frac{(\mu_p - p_1)}{\alpha\sigma_p^2}.$$ (26)

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 helpful interpretation is that agents are more sure about their signals. Let $\tau_p = \frac{1}{\sigma_p^2}$ and $\tau_{A|s+\xi} = \frac{1}{(\sigma_{A|s}^2 + \sigma_\xi^2)}$. We can rewrite the demand as

$$x_{2,A} = (\tau_{A|s+\xi} + \tau_p)\frac{\mu_A - p_1}{\alpha}$$ (27)

$$\equiv \tau_{p,d}\frac{\mu_A - p_1}{\alpha},$$ (28)

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

2.6.2 Inattentive agents

Inattentive agents' demand is given by

$$x_{2,I} = \frac{(\mu_I - p_1)}{\alpha(\sigma_{I|s}^2 + \sigma_\xi^2)} = \tau_{I|s+\xi}\frac{(\mu_I - p_1)}{\alpha}.$$ (29)

15

2.7 Market Clearing and Return Predictability

We have that the first-period price solves

$$V\pi_A x_{2,A} + V\pi_I x_{2,I} + u_1 = 0. \tag{30}$$

If we define weighted precisions as $\tau_{\pi_A} = \pi_A \tau_{p,d}$, $\tau_{\pi_I} = \pi_I \tau_{I|s+\xi}$, then we can express the price as

$$p_1 = \frac{\mu_A \pi_A \tau_{p,d} + \mu_I \pi_I \tau_I + \alpha \frac{u_1}{V}}{\pi_A \tau_{p,d} + \pi_I \tau_I} + = \frac{\mu_A \tau_{\pi_A} + \mu_I \tau_{\pi_I}}{\tau_{\pi_A} + \tau_{\pi_I}} + \frac{\alpha \frac{u_1}{V}}{\tau_{\pi_A} + \tau_{\pi_I}} \tag{31}$$

$$= \mu_E + \alpha \frac{u_1}{V} \sigma_E^2. \tag{32}$$

Where $\mu_E = \frac{\mu_A \tau_{\pi_A} + \mu_I \tau_{\pi_I}}{\tau_{\pi_A} + \tau_{\pi_I}}$ is then the economy-wide expectation, and $\sigma_E^2 = \frac{1}{\tau_{\pi_A} + \tau_{\pi_I}}$ is the economy-wide variance. The price is the dividend expectation weighted by each type of agent's (weighted) precisions plus the term corresponding to the nonfundamental demand divided by the volume of interest.

The immediate (unconditional) expected jump from the non-news price to the news price is

$$E[r_1] = E[p_1] - E[p_0] = \mu_E - \bar{d}, \tag{33}$$

which corresponds to the returns we would observe in high-frequency data. On average, the price change follows the direction of the news, albeit there can be deviations due to nonfundamental trading.

The expected price of period 2 is simply μ_A. Hence, the expected (dollar) returns, conditional on period 1, between period 2 and period 1 are given by

$$E_1[p_2 - p_1] = \mu_A - \mu_E + \alpha u_1 \sigma_E^2. \tag{34}$$

For example, if there is bad news ($s < \mu_A < \bar{d}$,) it will result, in underreaction ($\mu_A <$

$\mu_E < \bar{d}$). In that case, the price will be, on average, lower in the first period relative to the initial, and there will be some return predictability.

Proposition 2.1. *There is no unconditional return predictability.*

$$E[r] = E[p_2 - p_1] = 0 \tag{35}$$

Proof. In the Appendix. □

Intuitively, agents do not get positive or negative signals on average. Hence, the signal has the same expected value as the dividend, and the expected updated expectation is precisely the dividend for both the attentive and inattentive agents.

Proposition 2.2. *The return predictability from the point of view of attentive investors conditional on the signal being good is:*

1. *Positive:* $E_A[r|s > \bar{d}] = E_A[p_2 - p_1|s > \bar{d}] > 0$

2. *Decreasing in inattentive agent's capacity ω, the proportion of attentive agents π_A, the signal realization s, and the total volume V*

$$\frac{\partial E_A[r|s > \bar{d}]}{\partial \omega}, \frac{\partial E_A[r|s > \bar{d}]}{\partial \pi_A}, \frac{\partial E_A[r|s > \bar{d}]}{\partial s}, \frac{\partial E_A[r|s > \bar{d}]}{\partial V} < 0 \tag{36}$$

3. *Increasing in the risk aversion α and the noise trader variance σ_u^2*

$$\frac{\partial E_A[r|s > \bar{d}]}{\partial \alpha}, \frac{\partial E_A[r|s > \bar{d}]}{\partial \sigma_u^2} > 0 \tag{37}$$

4. *First increasing and then decreasing in attentive agents' capacity γ_A, the precision of the signal τ_S, and the precision of the dividend τ_d.*

With a symmetric effect for bad news.

Proof. In the Appendix. □

Intuitively, attentive agents comprehend that inattentive agents will eventually incorporate the entire information set. They do not entirely exploit this return predictability

17

since the price may not converge in their trading horizon due to nonfundamental demand. Consequently, attentive agents make profits, and inattentive agents have losses.

To determine the equilibrium proportion of informed agents, we assume there is a fixed cost of participating in this market as an attentive investor of $c_A \geq 0$. In that case, π_A adjusts until the profits of attentive agents equal $\tau_{p,d} \frac{(\mu_A - p_1)^2}{\alpha} = c_A$. We model c_A as an exogenous constant, higher when participating as an attentive agent that can short stocks. These costs stem from trading fees and investment restrictions, as only certain participants can short stocks.

Importantly, return predictability is an equilibrium outcome. Return predictability persists because it compensates agents for the cost of participating in the market as an attentive agent.

Proposition 2.3. *The return predictability is higher in smaller firms and with negative news.*

Proof. In the Appendix. □

2.8 Large Language Models

From the point of view of an LLM, the expectation about the dividend is given by

$$\mu_L \equiv E_L[d|s] = \frac{\bar{d}\tau_d + s\lambda\tau_S}{\tau_d + \lambda\tau_S}. \tag{38}$$

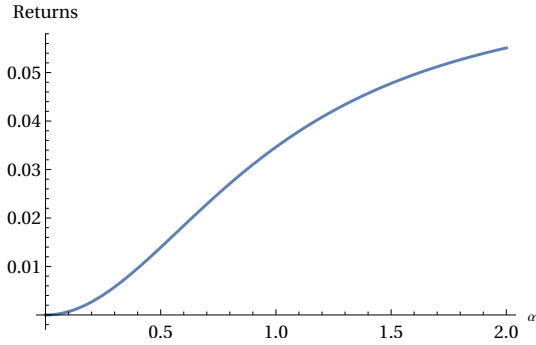And the updated precision by $\tau_{L|s} = \tau_d + \lambda\tau_S$.

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

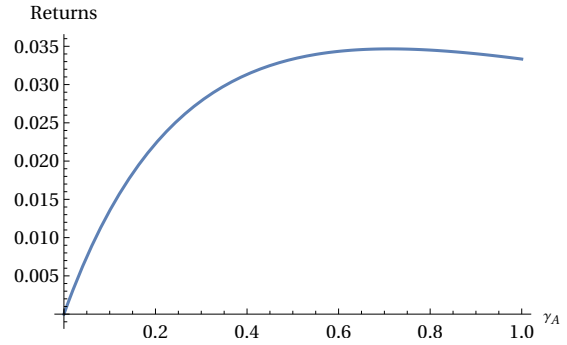$$x_{2,L} = \tau_{L|s}\frac{(\mu_L - p_1)}{\alpha} \tag{39}$$

Without considering price impact and from the point of view of the agent using the LLM, the strategy has an expected profit of

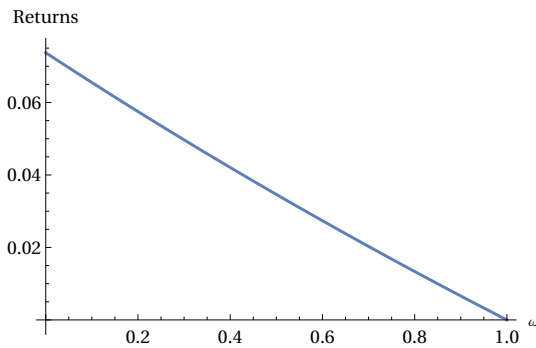$$x_{2,L}(\mu_L - p_1) = \tau_{L|s}\frac{(\mu_L - p_1)^2}{\alpha}. \tag{40}$$
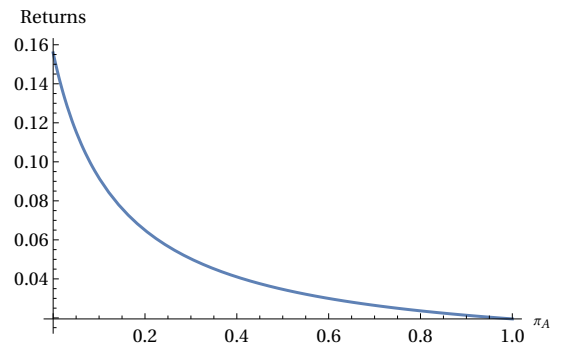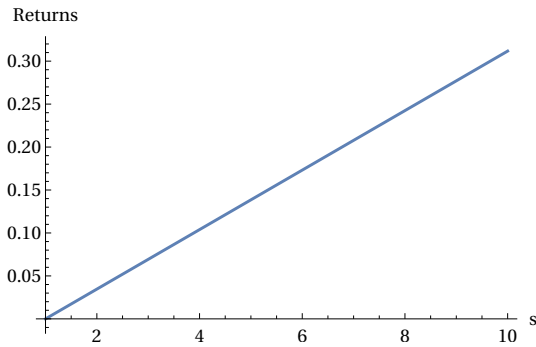
Figure 1: Returns without LLMs
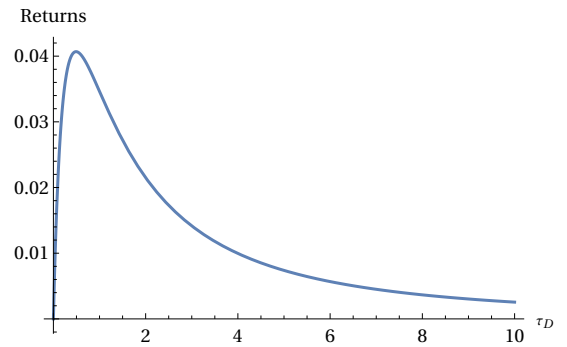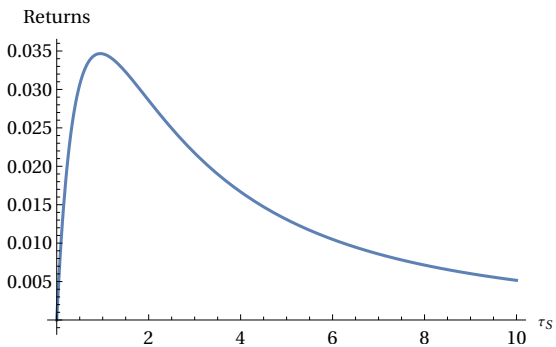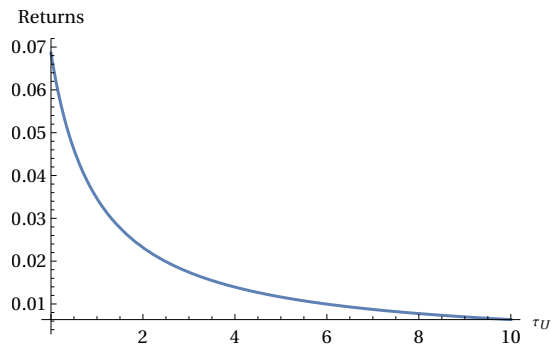
However, from the perspective of an attentive agent, this strategy has an expected profit
of

$$x_{2,L}(\mu_A - p_1) = \tau_{L|s} \frac{(\mu_L - p_1)(\mu_A - p_1)}{\alpha}. \tag{41}$$

The strategy can predict returns on average as long as $(\mu_L - p_1)$ has the same sign as
$(\mu_A - p_1)$.

Theorem 2.1. *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, the precision of the signal τ_S, and
   the precision of the dividend τ_d.*

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

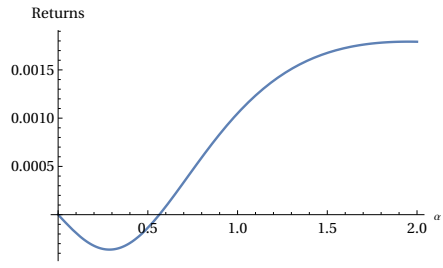*Proof.* In the Appendix. □

Proposition 2.4. *LLM-return predictability should be larger in markets with a lower pro-
portion of attentive market participants, such as smaller or illiquid stocks.*

However, the previous proposition abstracts the price impact associated with trading
based on LLMs' signals. Let's assume that inattentive agents use LLMs' signals instead
of their constrained capacity to trade. Then, by Proposition 2.3, return predictability will
decline as markets become more efficient. Moreover, all things equal, we would expect return
predictability to decline more with larger-size LLMs.

Let's first consider the case where a fraction of inattentive agents, θ, starts processing
the information with LLMs. We still assume the remaining information will be processed in
the second period. Then the new price in the first period, $p_{1,L}$ is given by

$$p_{1,L} = \frac{\mu_A \pi_A \tau_{p,d} + (1-\theta)\mu_I \pi_I \tau_I + +\theta \pi_I \mu_L \tau_{L|s}}{\pi_A \tau_{p,d} + (1-\theta)\pi_I \tau_I + \theta \pi_I \tau_{L|s}} \tag{42}$$

20

Figure 2: Return Predictability Forecasting With LLMs

(a)
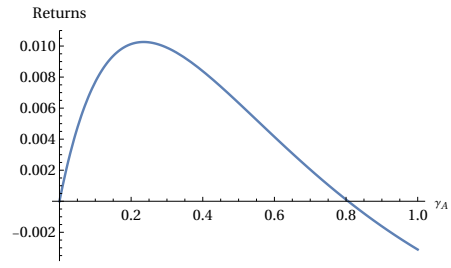
(b)

(c)

(d)

(e)

(f)
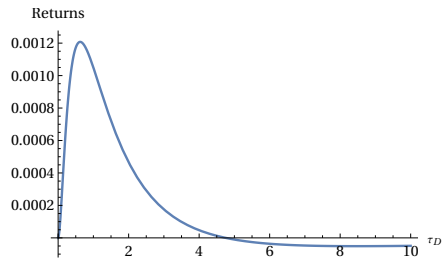
(g)
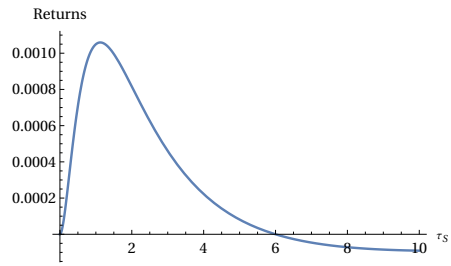
(h)
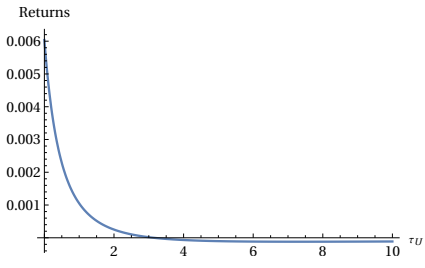
(i)

Theorem 2.2. *Price informativeness increases in the proportion of agents using LLMs, and on LLMs model size k. The price is more informative than in the equilibrium without LLMs as long as $\lambda(c, k) > \omega\gamma_A$.*

Proof. In the Appendix. □

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 information.

Theorem 2.3. *If LLMs have a sufficiently large model size and all inattentive agents use them, there will no longer be any return predictability.*
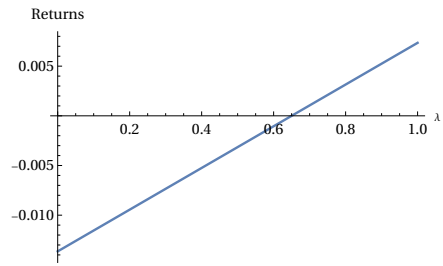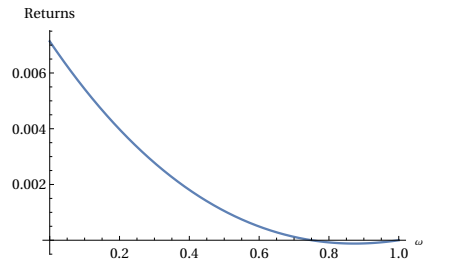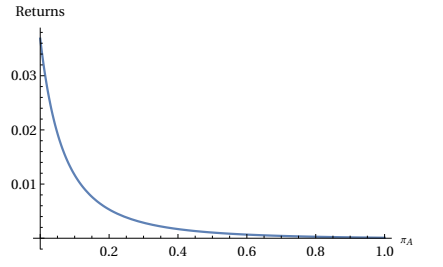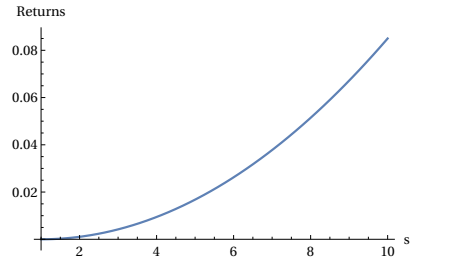
Proof. In the Appendix. □

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.

2.9 LLMs Better than Attentive Agents

We will focus on the case where LLMs become better than attentive agents at processing information, and all attentive agents switch to the better LLM technology. Implicitly, we assume the technology is widely available, as is the case for the current best LLMs.

That case is analogous to increasing attentive agents' information capacity, γ_A, and Proposition 2.3 applies.

Proposition 2.5. *There is a unique value γ_A^* (keeping other parameters fixed) such that for $\lambda c, k < \gamma_A^*$, increasing $\lambda c, k$ increases the return predictability. In contrast, if $\lambda c, k > \gamma_A^*$, increasing attentive agents' information capacity decreases the return predictability.*

Proposition 2.6. *If attentive agents use better-than-human LLMs, price informativeness increases in LLMs' model size.*

The more complex case occurs when only a fraction of attentive agents switch. 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 signal that attentive agents that use the LLMs process. 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 of autonomous LLMs trading, and the implications are similar to the above cases.

3   Data

We utilize three primary datasets for our analysis: the Center for Research in Security Prices (CRSP) daily returns, news headlines, and RavenPack. The sample period begins in October 2021 (as ChatGPT's training data is available only until September 2021) and ends in December 2022. This sample period ensures that our evaluation is based on information not present in the model's training data, allowing for a more accurate out-of-sample assessment of its predictive capabilities.

The CRSP daily returns dataset contains information on daily stock returns for a wide range of companies listed on major U.S. stock exchanges, including data on stock prices, trading volumes, and market capitalization. This comprehensive dataset enables us to examine the relationship between the sentiment scores generated by ChatGPT and the corresponding stock market returns. 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.

We first collect a comprehensive news dataset for all CRSP companies using web scraping. We search for all the news containing the company name or the ticker. The resulting dataset comprises news headlines from various sources, such as major news agencies, financial news websites, and social media platforms. For each company, we collect all the news during the sample period. We then match the headlines with those from a prominent news

sentiment analysis data provider (RavenPack). We match the period and the news title for all companies that have returns on the following market opening. Most (more than 70%) of matched headlines correspond to press releases. We do not use the RavenPack enhance headlines that potentially contain more information since they are not widely disseminated to the public. We match 67,586 headlines of 4,138 unique companies. We process the merged dataset using the preprocessing methods outlined by Jiang, Li, and Wang (2021).

Importantly, 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 may have come from the web scraping procedure. Further, we employ their news categorization to explain the differences in return predictability across different models. Moreover, they have a close mapping with CRSP, which ensures the news is matched with returns at the exact time. We further use their infrastructure by using only the information they consider highly relevant for a given company in a given period.

We employ the "relevance score" from the data vendor, which ranges from 0 to 100, 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. These filters are imposed to ensure that we analyze fresh and relevant news headlines that could impact stock prices, which provides power to test the capabilities of LLMs in predicting stock market movements.

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 headlines with a similarity greater than 0.6 for the same company on the same day. These filtering techniques do not introduce look-ahead bias, as the data vendor evaluates all news articles within milliseconds of receipt and promptly sends the resulting data to users. Consequently, all information is

available at the time of news release.

Table 1 presents selected descriptive statistics of the sample: (i) the daily stock returns (%), (ii) the headline length, (iii) the response length, (iv) the GPT score (1 if ChatGPT 3.5 says YES, 0 if UNKNOWN, and -1 if NO), and the event sentiment score provided by the data vendor. The average GPT score is positive (0.24), with the median being zero, and the event sentiment score shows a similar pattern. Thus, news headlines have an overall positive tilt. Panel B reports the correlation matrix of these variables. The correlation between the GPT score and event sentiment score is low, less than 0.28.

4 Methods

4.1 Prompt

Prompts are critical in guiding ChatGPT's responses to specific tasks and queries. A prompt is a short text that provides context and instructions for ChatGPT to generate 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.

The prompt serves as the starting point for ChatGPT's response generation process. The model uses the information contained in the prompt to generate a relevant and contextually appropriate response. This process involves analyzing the syntax and semantics of the prompt, developing a series of possible answers, and selecting the most appropriate one based on various factors, such as coherence, relevance, and grammatical correctness.

Prompts are essential for enabling 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 and generate responses tailored to the user's needs. Moreover, prompts can be customized to perform tasks in different domains, such as finance, healthcare, or customer support.

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 _term_ term?

Headline: _headline_

In this prompt, we ask ChatGPT, a language model, 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. _term_ corresponds to either short or long-term. 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 implicitly assumed that the headline contains sufficient information for an expert in the financial industry to reasonably assess its impact on the stock price. This prompt is designed to demonstrate the capabilities of ChatGPT as a language model in financial analysis tasks. We set the temperature of GPT models to 0 to maximize the reproducibility of the results.[7]

For example, consider the following headline about Oracle:

Rimini Street Fined $630,000 in Case Against Oracle.

The prompt then asks:

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 Oracle in the short term?

Headline: Rimini Street Fined $630,000 in Case Against Oracle

---

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

And here is ChatGPT's response:

YES

The fine against Rimini Street could potentially boost investor confidence in Oracle's ability to protect its intellectual property and increase demand for its products and services.

The news headline states that Rimini Street has been fined \$630,000 in a case against Oracle. The proprietary software analytics tool gives a negative sentiment score of -0.52, indicating that the news is perceived as unfavorable. However, ChatGPT responds that it believes the information to be positive for Oracle. ChatGPT reasons that the fine could increase investor confidence in Oracle's ability to protect its intellectual property, potentially leading to increased demand for its products and services. This difference in sentiment highlights the importance of context in natural language processing and the need to carefully consider the implications of news headlines before making investment decisions.

4.2 Empirical Design

We prompt ChatGPT to provide a recommendation for each headline and transform it into a numerical "ChatGPT score," where "YES" is mapped to 1, "UNKNOWN" to 0, and "NO" to -1. We average the scores if there are multiple headlines for a company on a given day. We match the headlines to the next trading period. For headlines before 6 a.m. on a trading day, we assume the headlines can be traded by the market opening of the same day and sold at the close of the same day. For headlines after 6 a.m. but before 4 p.m., we assume the headlines can be traded at the same day's close and sold at the close of the next trading day. For headlines after 4 p.m., we assume the headlines can be traded at the opening price of the next day and sold at the closing price of that next day. We then run linear regressions of the next day's stock returns on the ChatGPT score, the sentiment score provided by the data vendor, and scores from other LLMs. Thus, all of our results are out-of-sample.

Specifically, we estimate the following regression specification:

$$r_{i,t+1} = a_i + b_t + \gamma' x_{i,t} + \varepsilon_{i,t+1}, \tag{43}$$

27

where the dependent variable, $r_{i,t+1}$, is stock i's return over a subsequent trading day as discussed above, $x_{i,t}$ refers to the vector containing the ChatGPT score or other scores 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.

In addition to analyzing the performance of ChatGPT, we examine the capabilities of other more basic models, such as BERT, GPT-1, and GPT-2, and compare their performance 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 B provides more details on the prompts we use for these models. Contrasting the performance of the basic and the more advanced LLMs allows us to shed light on whether return predictability is an emerging capacity of the recent developments in language models.

5 Can ChatGPT Predict Stock Returns?

5.1 Results from Long-Short Trading Strategies

To assess ChatGPT's capabilities to predict stock price movements, we examine the performance of long-short strategies formed based on ChatGPT scores of news headlines. In particular, we form zero-cost portfolios that buy the stocks with a positive ChatGPT score and sell stocks with a negative ChatGPT score after the news release. If a piece of news is released before 6 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 released after 6 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. 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.

Figure 1 plots the cumulative returns of seven different trading strategies (investing

$1) without considering transaction costs. These seven strategies include (1) an equal-weighted portfolio that buys companies with good news based on ChatGPT 3.5 ("Long"), (2) an equal-weighted portfolio that sells companies with bad news based on ChatGPT 3.5 ("Short"), (3) a self-financing long-short strategy based on ChatGPT 3.5 ("Long - Short"), (4) a self-financing long-short strategy based on ChatGPT 4 ("Long - Short GPT 4"), (5) an equal-weight market portfolio (Market Equally-Weighted), (6) a value-weight market portfolio ("Market Value-Weighted"), and (7) an equal-weight portfolio in all stocks with news the day before regardless of the news direction ("All News").

We find strong evidence of the power of ChatGPT scores in predicting stock returns the next day. For instance, without considering transaction costs, a self-financing strategy that buys stocks with a positive ChatGPT-3.5 score and sells stocks with a negative ChatGPT-3.5 score earns an astonishing cumulative return of over 550% from 2021m10 – 2022m12. In contrast, the equal-weight and value-weight market portfolios and a naive equal-weight all-news portfolio all earn negative cumulative returns over the same periods. The sharp difference in performance across these two sets of portfolios demonstrates that as a leading LLM, ChatGPT can add value by extracting valuable information from news headlines and predicting stock market reactions. We note that both the long and short legs contribute to the predictability of ChatGPT. While the long leg delivers about 200%, the short leg delivers over 250%. Thus, the predictability is even stronger among stocks with bad news.

One may argue that the analysis in Figure 1 ignores transaction costs and may or may not hold after incorporating transition costs. In Figure 2, we evaluate the performance of the long-short strategy based on ChatGPT 3.5 under different transaction cost assumptions: 5, 10, and 25 basis points (bps) per transaction. Even assuming a transaction cost of 5 bps per trade (i.e., 10 bps round-trip), the strategy still earns a cumulative return of over 450% during our sample period. As we increase the transaction costs to 10 bps per trade, the cumulative return is still as high as 350%. Assuming a very high transaction cost of 25 bps per trade, the cumulative return falls to 50%. As the short leg tends to deliver higher cumulative returns, we evaluate its performance under different transaction cost assumptions in Figure 3. We find that the cumulative return of the short leg is more sensitive to transaction cost changes than the long-short portfolios. Increasing the transaction costs from 0 to 25 bps per trade

will erode all the positive 250%+ cumulative returns and make it negative, which is on par with the equal-weighted market portfolio.

Turning our attention to the more complex ChatGPT 4 model, the long-short strategy generates a cumulative return of over 350%, which also clearly outperforms the market portfolios or the naive all-news portfolio. While ChatGPT 4's cumulative return is lower in magnitude compared to that of the long-short strategy based on ChatGPT 3.5, it has much lower variations over time.

In Table 2, we show the Sharpe ratio and maximum drawdown of the seven strategies shown in Figure 1. We find that the strategy based on ChatGPT 4 delivers a much higher Sharpe ratio of 3.8, compared to a Sharpe ratio of 3.1 for the one based on ChatGPT 3.5. In addition, the ChatGPT 4 strategy has a maximum drawdown of -10.4% compared to the maximum drawdown of -22.8% for the ChatGPT 3.5 strategy. Thus, it appears that the predictability is better achieved by the more complex models like ChatGPT-4.

5.2 Results from Predictive Regressions

In this section, we use prediction regressions to evaluate the performance of different LLMs. Specifically, we carry out regressions as in Equation 43, using scores from different LLMs from assessing news headlines to predict next-day stock returns. Firm and time-fixed effects are included in the regressions, and standard errors are clustered by date and firm.

We present the results from regressions of stock returns on prediction scores from more advanced LLMs in Table 3. The main variables of interest include scores from three advanced LLMs: (i) ChatGPT 3.5, (ii) ChatGPT 4, and (iii) BART Large. First, we find that the prediction score from ChatGPT 3.5 has a statistically and economically significant relation with the next-day stock returns. Specifically, the coefficient on ChatGPT 3.5's score is 0.259 with a *t*-stat of 5.259. A switch from a negative (-1) to a positive (1) prediction score is associated with a 51.8 bps increase in next-day stock return. This result highlights 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. In our analysis, we control for the ChatGPT sentiment

30

scores and examine the predictive power of these alternative sentiment measures. Our results show that when controlling for the ChatGPT sentiment scores, the effect of the sentiment score from the data vendor on daily stock market returns is attenuated. This indicates that the ChatGPT model outperforms existing sentiment analysis methods in forecasting stock market returns.

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. These findings confirm the predictive power of ChatGPT sentiment scores and emphasize the potential benefits of incorporating LLMs into investment decision-making processes. By outperforming traditional sentiment analysis methods, ChatGPT demonstrates its value in enhancing the performance of quantitative trading strategies and providing a more accurate understanding of market dynamics.

Third, our analysis reveals that ChatGPT 4 sentiment scores exhibit a strong and positive significant predictive power on daily stock market returns. Consistent with the evidence in Figure 1 and Table 2, the coefficient on the ChatGPT 4 score is lower than that of the ChatGPT 3.5 score, but the former has a large t-stat. Again, when we include both scores from ChatGPT 4 and the data vendor in the same regressions, only the coefficient on ChatGPT 4's score is positive and significant, while the coefficient on the data vendor's score is not.

To compare various language models' performance, we conduct a similar regression analysis using prediction scores from other LLMs. In particular, we consider six more basic LLMs in Table 4: (i) DistilBart-MNLI-12-1, (ii) GPT-2 Large, (iii) GPT-2, (iv) GPT-1, (v) BERT, and (vi) BERT Large. Our results show a striking pattern—return predictability is an emerging capacity of more complex language models. Scores from BART Large and DistilBart-MNLI models show some predictability but are noticeably weaker compared to ChatGPT 3.5 and 4. When we use more 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. In contrast, we observe the highest predictability in the most

complex model—ChatGPT-4.

In the next set of analyses, we examine the predictability of our complete list of LLMs across small and large-cap stocks in Tables 5-8. Our results show that the predictability of the ChatGPT scores is present among both small and large-cap stocks. This suggests that the market appears to be underreacting to firm-specific news at the daily frequency we examine, consistent with the evidence in the long literature (e.g., Bernard and Thomas (1989), Chan, Jegadeesh, and Lakonishok (1996), DellaVigna and Pollet (2009), Hirshleifer, Lim, and Teoh (2009), and Jiang, Li, and Wang (2021)). Nevertheless, we also find that the predictability is more pronounced among smaller stocks. For instance, the coefficient on the ChatGPT 3.5 score in predicting returns of small stocks (i.e., less than the 10th percentile NYSE market cap distribution) in Table 5 is more than four times the magnitude of the one for the remaining sample in Table 7 (0.653 with a t-stat of 5.145 vs 0.148 with a t-stat of 3.084). This observation is consistent with the idea that limits-to-arbitrage also play an essential role in driving this predictability.

5.3 Additional Analyses

5.3.1 Large vs. Small Cap Stocks

In this section, we carry out more analysis to further understand the capabilities of the different LLMs in predicting stock market returns. Table 8 reports the average returns and abnormal returns based on the CAPM model and the 5-factor model of Fama and French (2015) for the different models. Panel A analyzes the full sample of common stocks, Panel B analyzes small stocks below the 10th percentile NYSE market capitalization, and Panel C analyzes the remaining non-small stocks. We find consistent evidence that more complex models are generally better. The magnitude and t-stats of average returns and alphas are significantly higher for more advanced models like GPT-4 and GPT-3.5. For instance, the average daily return and the 5-factor alpha for GPT-4 are 44 bps (t-stat.=4.24) and 41 bps (t-stat.=4.01), respectively, both economically and statistically significant. In contrast, more basic models like GPT-1, GPT-2, and BERT do not generate significantly positive returns. This pattern holds for both small and large stocks.

Furthermore, Table 9 reports the Sharpe ratios and number of stocks in each leg by different models. As in the previous table, Panel A analyzes the full sample of common stocks, Panel B analyzes small stocks below the 10th percentile NYSE market capitalization, and Panel C analyzes the remaining non-small stocks. While the long-short strategy based on GPT-3.5 generates higher average returns and alphas in magnitude, it holds less diversified positions. For instance, as shown in Panel A, the average numbers of stocks in the short and long legs of the GPT-3.5 strategy are 3.8 and 50.9, respectively. The corresponding numbers for the GPT-4 strategy are 23.3 and 98.5, respectively, significantly higher than those for the GPT-3.5 strategy, especially for the short leg. As a result of a more diversified portfolio in both legs, the Sharpe ratio of the strategy based on GPT-4 is 3.8, which is higher than 3.1 of GPT-3.5. We find a consistent pattern that more complex models have higher Sharpe ratios. Notably, the complex models continue to have a high Sharpe ratio in non-small stocks. For instance, as shown in Panel C, the Sharpe ratio for GPT-4 is as high as 2.99, even after removing the small stocks. Overall, the general ability of models to understand natural language appears to be positively correlated with their ability to forecast returns accurately.

5.3.2 News Complexity

We further examine the predictive capabilities of various language models on stock market returns and the complexity of news headlines. As detailed in Table 10, 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 more complicated to understand. The table reports the annualized Sharpe ratio of long-short portfolios implied by these models.

Table 10 provides a clear differentiation in model performance. Advanced models like GPT-3.5 and GPT-4 demonstrate a higher capability in interpreting complex headlines, as seen in their similar or slightly higher Sharpe ratios for high-complexity news compared to low-complexity news. On the other hand, basic models like GPT-1, GPT-2, Bert, and Fin-bert show limited forecasting abilities in high-complexity news, while many of these models have some capabilities of interpreting low-complexity news. This suggests that the more so-

phisticated models have distinct advantages in processing detailed information from complex news sources. The efficiency with which advanced models process information showcases the potential evolution of information dissemination in computational systems. The emergence of these advanced models has tangible implications for the labor market. As they become more proficient at deciphering and interpreting multifaceted information, there's a potential shift in the labor dynamics. Traditional roles that depended on human interpretation could evolve, leading to a surge in demand for specialists capable of developing, training, and overseeing these models. This shift could notably influence labor trends, especially in the finance and tech sectors.

5.3.3   News Types

Firm news could come from different sources, with some from news media and others directly released by firms. Table 11 delineates the annualized Sharpe ratios derived from various models, indicating their predictive efficiency based on different sources of information. The table distinguishes between the outcomes when using news articles vs. press releases. The more sophisticated AI models, specifically GPT-4, outshine others with higher Sharpe ratios across both categories, implying a robust ability to interpret and utilize information from diverse news sources. Notably, GPT-4's performance remains superior even when analyzing press releases, which could be more susceptible to corporate influence given their usual origin from within firms. This trend of higher ratios persists with other advanced models like GPT-3.5 and Distilbart-Mnli-12-1, especially when dealing with news articles, which are third-party communications and may present a less biased view than press releases. The contrast is stark when compared to simpler models like GPT-1, BERT, and FinBERT, which show diminished predictive power, particularly with press releases, sometimes yielding negative Sharpe ratios, suggesting a loss rather than a gain from their predictions in these contexts. The ability of complex models to decipher the subtleties in press releases highlights their resilience in the presence of nuanced signals that the issuing firms could potentially manipulate.

6 Interpretability

6.1 Evaluating ChatGPT's Reasoning Capabilities

Traditional machine learning models in finance primarily focus on prediction, often lacking interpretability. In contrast, large language models (LLMs) like ChatGPT offer predictions and associated explanations in natural language. This distinctive feature provides a deeper insight into the rationale behind each prediction, a capability largely absent in conventional models. Motivated by this unique attribute, we devise a novel framework to harness these qualitative insights for enhanced predictive accuracy. In our exploration of ChatGPT's capabilities, every headline processed by the model yielded a prediction alongside an explanation. Unlike traditional quantitative models offering limited transparency, ChatGPT presents contextual reasoning in plain language, adding a rich layer of qualitative information.

Building on this capability, we analyze ChatGPT-4's predictions more deeply, emphasizing the qualitative explanations accompanying each forecast. We postulate that these explanations, filled with nuanced details, could be instrumental in return predictability. By examining these textual rationales, we hypothesize that they contain critical information that can significantly enhance the accuracy of financial forecasts.

The initial phase of our process centers on refining ChatGPT-4's textual explanations to distill the core reasoning. We programmatically extract the raw explanation text accompanying each prediction, ensuring the separation of these qualitative statements from the primary forecasts. Subsequent cleaning involved removing explicit sentiment indicators like 'YES,' 'NO,' and 'UNKNOWN,' which, while useful for direct predictions, detracts from the sentiment's qualitative depth. Our primary objective is to focus on the essence of the model's rationale, devoid of overt prediction markers, setting the stage for a thorough analysis.

Building on this foundation, we aim to translate ChatGPT's explanations into a more quantifiable format. We exclude neutral scores and segment the reasons based on positive or negative recommendations to achieve this. This segmentation allows for a more focused analysis, underpinning our hypothesis that the correctness of an explanation is different if it deals with optimistic or pessimistic predictions.

Armed with the refined explanations, our primary challenge is mutating this qualitative

data into structured quantitative data for regression analysis. To achieve this transformation, we turn to the Term Frequency-Inverse Document Frequency (TF-IDF) technique, a standard method in natural language processing. TF-IDF, by design, weights words based on their prevalence in individual documents while discounting their frequency across the entire corpus. This ensures common terms receive lower weights, whereas distinctive words — potentially indicative of specific ideas — are emphasized. In the context of our research, such emphasis on unique terms is invaluable, as it spotlights words potentially expressive of strong market sentiments and potential stock movements.

Figures **??**, **??**, **??**, and **??** show the words with the highest overall TF-IDF scores by explanation for all explanations, positive, negative, and neutral, respectively. For all explanations, there is no particular pattern with words such as 'stock,' 'price,' 'impact,' 'results,' 'depends,' 'financial,' 'company,' 'specific,' 'reported,' 'confidence,' 'performance,' 'indicates,' 'short,' 'term,' 'potentially,' 'shares,' 'conference,' 'details,' 'actual,' and 'indicate' appear.[8]

For positive explanations, we start looking at words indicative of positive outcomes and sound financial performance with words such as 'dictates,' 'positive,' 'company,' 'stock,' 'performance,' 'price,' 'confidence,' 'strong,' 'increased,' 'financial,' 'dividend,' 'growth,' 'likely,' 'future,' 'investors,' 'revenue,' 'positively,' 'typically,' 'boost,' and 'potential' appearing. For neutral explanations, we have only general descriptive indicators, financial metrics, and uncertainty or conditionality with words such as 'impact,' 'results,' 'price,' 'stock,' 'depends,' 'specific,' 'reported,' 'financial,' 'company,' 'conference,' 'confidence,' 'actual,' 'details,' 'shares,' 'potentially,' 'depend,' 'short,' 'term,' 'performance,' and 'lack.' Finally, for negative explanations, we have words indicating adverse outcomes, financial concerns and legal issues such as 'negatively,' 'term,' 'short,' 'stock,' 'price,' 'impact,' 'lawsuit,' 'action,' 'class,' 'negative,' 'company,' 'potential,' 'impacted,' 'indicates,' 'lower,' 'sales,' 'indicate,' 'securities,' 'outlook,' and 'financial.' However, this method does not allow distinguishing between correct and incorrect explanations. To do so, we use a supervised approach to characterize the words that best predict if a given answer relates correctly to the outcome.

---

8. In Figures **??** and **??**, we also present the length distribution for all headlines and all explanations, respectively. Figure 6g further shows the average Cosine similarity of all explanations, with the vast majority of the density falling in the range of 0.1 and 0.3.

Having transformed the textual explanations into a structured format, we employ regularized logistic regression models. The choice of logistic regression is motivated by our binary outcome variable: whether the stock price moved in the direction predicted by ChatGPT. By modeling this relationship, we aim to establish the predictive power of the qualitative explanations provided by the model. By training distinct models for positive and negative news explanations, we aim to capture the subtle variations in how optimistic and pessimistic rationale influenced stock price dynamics. This approach provides a more comprehensive view of the market's sentiment-driven behavior.

In our analysis, understanding the significance and influence of individual features, particularly words in this context, is paramount. To achieve this, we first extract the terms with the highest and lowest coefficients in the logistic regression models. This method provides positive and negative coefficients, allowing us to identify words that had the most pronounced impact on whether the explanations were correct. Each set of words is categorized into "correct" and "incorrect" based on their effect on the predictive accuracy. These words could shed light on the lexicon that ChatGPT relied on and illustrate how certain terms were more aligned with specific ideas.

While individual words are interesting, it is difficult to understand the full impact without their context. To delve deeper into the contextual environment of these influential words, we identify the words that frequently accompany a given target word. We filter out less significant words by setting a threshold percentile for the average TF-IDF, ensuring that the accompanying words list was both relevant and impactful. Through this approach, we are not just looking at words in isolation but also understanding their significance in the broader context, shedding light on the intricate web of relationships that underpin stock market predictions.

6.2   Interpretability Results

Table 12 reports the exercise results for ChatGPT-4's positive recommendations with several themes. First, the model predicts satisfactorily when its reasoning is related to stock purchases by insiders. Second, the model forecasts accurately when its explanations relate to earnings guidance. Third, the model performs well for themes related to earnings per

share or market share. Finally, the model also does well when the theme relates to dividends. On the contrary, the model does worse when its reasoning refers to partnerships or new developments. It also fails when justifying the recommendation with profits, sales, and profitability. One potential reason for the negative influence of profits, sales, and profitability could be that ChatGPT was fed with only the news headlines but not the market expectation regarding firms' profits or sales at the time of news release. For scheduled events like earnings releases, it is crucial to use the market expectation as the benchmark to tease out the component that moves the market.

Table 13 reports the exercise results for ChatGPT-4's negative recommendations with several themes. First, the model predicts well when its reasoning is related to risk of downgrade or risk related to credit. Second, the model also predicts satisfactorily when the theme is related to factors that impacted earnings or revenue negatively. Third, the model forecasts accurately when the theme is related to fraud or reputational damages. Finally, the model also does well when its explanations relate to the sale of securities by directors. On the contrary, the model does worse when its reasoning relates to prospects or outlook. It also fails when reasoning about profits, sales, and profitability.

7 Conclusion

In this study, we investigate the potential of ChatGPT and other large language models in predicting stock market returns using sentiment analysis of news headlines. We document several findings that are 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 from a leading data vendor. Second, 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. Third, the predictability of ChatGPT scores presents in both small and large stocks, suggesting market underreaction to company news. Fourth, predictability is stronger among smaller stocks and stocks with bad news, consistent with limits-to-arbitrage playing an important role. Finally, we

propose a new method to evaluate and understand the models' reasoning capabilities. By demonstrating the value of LLMs in financial economics, we contribute 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 market returns, researchers can develop more targeted strategies for improving these models and maximizing their utility in finance.

Additionally, 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. Future research can explore the role of LLMs in shaping market behavior and their potential positive and negative consequences for the financial system.

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.

In short, our study demonstrates the value of ChatGPT in predicting stock market returns. It paves the way for future research on the applications and implications of LLMs in the financial industry. As the field of AI-driven finance continues to expand, the insights gleaned from this research can help guide the development of more accurate, efficient, and responsible models that enhance the performance of financial decision-making processes.

References

Acemoglu, Daron, David Autor, Jonathon Hazell, and Pascual Restrepo. 2022. "Artificial Intelligence and Jobs: Evidence from Online Vacancies." *Journal of Labor Economics* 40, no. S1 (April): S293–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 Xi He, and James Hodson. 2022. "Artificial Intelligence, Firm Growth, and Product Innovation." *SSRN Electronic Journal* (May).

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.

Binsbergen, Jules H van, Xiao Han, and Alejandro Lopez-Lira. 2023. "Man versus Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases." *The Review of Financial Studies* 36, no. 6 (May): 2361–2396.

Binsbergen, Jules H. van, Xiao Han, Alejandro Lopez-Lira, Jules H van Binsbergen, Xiao Han, and Alejandro Lopez-Lira. 2020. *Man vs. Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases.* Technical report, Working Paper Series 27843. National Bureau of Economic Research.

Bybee, Leland, Bryan T. Kelly, Asaf Manela, and Dacheng Xiu. 2019. "The Structure of Economic News." *Working Paper* (January).

———. 2021. "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.

Campbell, John L., Hsinchun Chen, Dan S. Dhaliwal, Hsin-min min Lu, Logan B. Steele, John L. Campbell, Hsinchun Chen, et al. 2014. "The information content of mandatory risk factor disclosures in corporate filings." *Review of accounting studies* (Boston) 19, no. 1 (March): 396–455.

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.

Chin, Andrew, and Yuyu Fan. 2023. "Leveraging Text Mining to Extract Insights from Earnings Call Transcripts." *Journal Of Investment Management* 21 (1).

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

Cowen, Tyler, and Alexander T. Tabarrok. 2023. "How to Learn and Teach Economics with Large Language Models, Including GPT." *SSRN Electronic Journal* (March).

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

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.

Gaulin, Maclean Peter. 2017. "Risk Fact or Fiction: The Information Content of Risk Factor Disclosures."

Gromb, Denis, and Dimitri Vayanos. 2010. "Limits of Arbitrage." *Annual Review of Financial Economics* 2 (November): 251–275.

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. 2022. "Surviving the FinTech disruption." *Working Paper.*

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.*

Korinek, Anton. 2023. "Language Models and Cognitive Automation for Economic Research." (Cambridge, MA) (February).

Lopez-Lira, Alejandro. 2019. "Risk Factors That Matter: Textual Analysis of Risk Disclosures for the Cross-Section of Returns." *SSRN Electronic Journal* (September).

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

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.

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.

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.

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).

Yang, Kai-Cheng, and Filippo Menczer. 2023. "Large language models can rate news outlet credibility." *Working Paper.*

Figure 3: Cumulative Returns of Investing $1 (Without Transaction Costs)
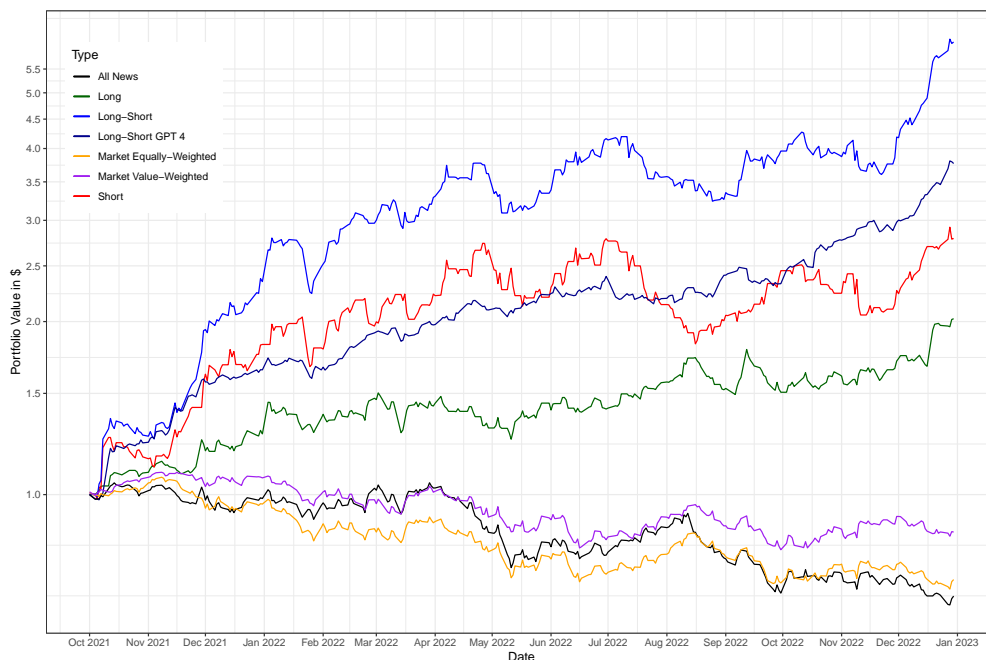
This figure presents the results of different trading strategies based on ChatGPT 3.5 and ChatGPT 4 without considering transaction costs. If a piece of news is released before 6 a.m. on a trading day, we enter the position at the market opening and exit at the close of the same day. If the news is released after 6 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. 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 "All-news" black line corresponds to an equal-weight portfolio in all companies with news the day before (regardless of news direction). The green line corresponds to an equal-weighted portfolio that buys companies with good news, according to ChatGPT 3.5. The red line corresponds to an equal-weighted portfolio that short-sells companies with bad news, according to ChatGPT 3.5. The light 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 3.5. The dark 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 yellow line corresponds to an equally weighted market portfolio. The purple line corresponds to a value-weighted market portfolio.

Figure 4: Cumulative Returns of Investing \$1 in the Long-Short Strategy for Different Transaction Costs

This figure presents the results of different trading strategies for different transaction costs. If a piece of news is released before 6 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 released after 6 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. 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 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 3.5, with zero transaction costs. The dark green line corresponds to the same equal-weighted zero-cost portfolio with a cost of 5 bps per transaction (i.e., 10 bps round trip). The light blue line corresponds to the same equal-weighted zero-cost portfolio with a cost of 10 bps per transaction. The dark blue line corresponds to the same equal-weighted zero-cost portfolio with a cost of 25 bps per transaction. The red line corresponds to an equally weighted market portfolio without transaction costs.

44

Figure 5: Cumulative Returns of Investing $1 in the Short Strategy for Different Transaction Costs
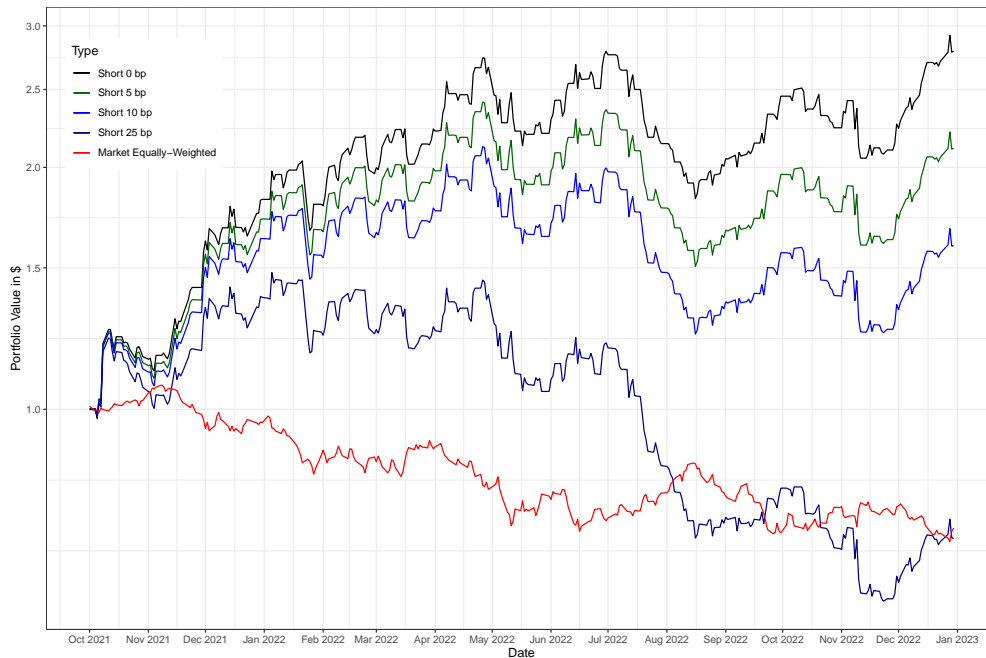
This figure presents the results of different trading strategies for different transaction costs. If a piece of news is released before 6 a.m. on a trading day, we enter the position at the market opening and exit at the close of the same day. If the news is released after 6 a.m. but before the market close, we enter the position at the market close price of the same day and exit at the close of the next trading day. 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 black line corresponds to an equal-weighted short portfolio that buys companies with good news and short-sells companies with bad news, according to ChatGPT 3.5, with zero transaction costs. The dark green line corresponds to the same equal-weighted short portfolio with a cost of 5 basis points per transaction. The light blue line corresponds to the same equal-weighted short portfolio with a cost of 10 basis points per transaction. The dark blue line corresponds to the same equal-weighted short portfolio with a cost of 25 basis points per transaction. The red line corresponds to an equally weighted market portfolio without transaction costs.

45
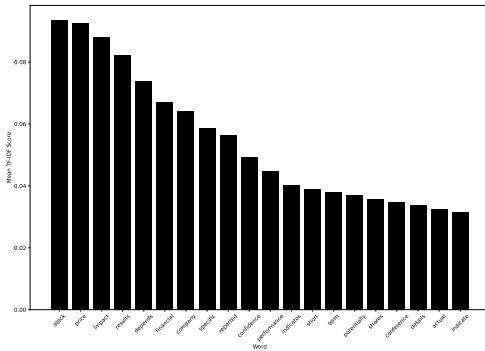
(a) Words with Highest TF-IDF Score for All Explanations
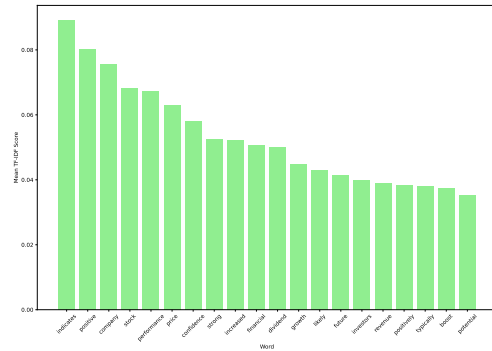
(b) Words with Highest TF-IDF Score for Positive Explanations
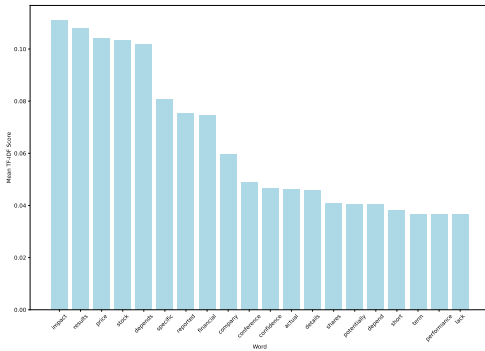
(c) Words with Highest TF-IDF Score for Neutral Explanations
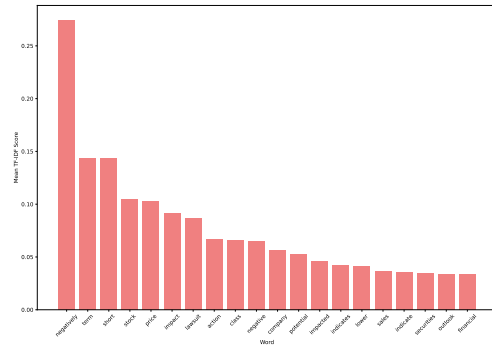
(d) Words with Highest TF-IDF Score for Negative Explanations
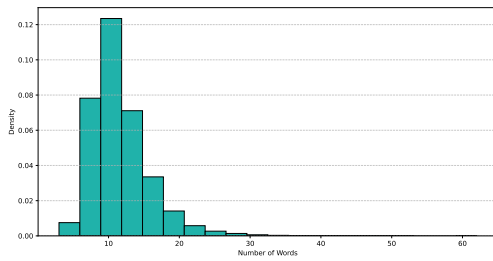
(e) Length Distribution for All Headlines
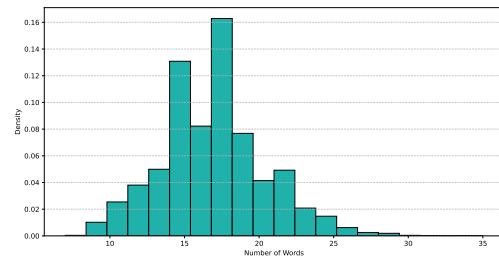
(f) Length Distribution for All Explanations
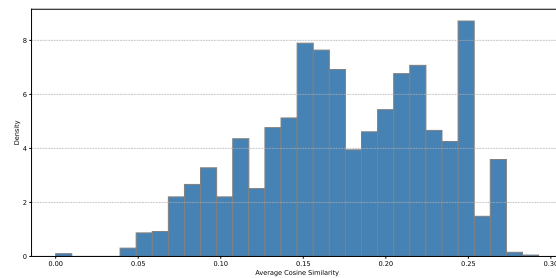
(g) Distribution of Average Explanation Similarity

Figure 6: Text Analysis Results

Table 1: Descriptive Statistics of Various Portfolios

This table reports the following statistics of the different trading strategies as specified in Figure 1: Sharpe ratio, mean daily returns, standard deviation of daily returns, and maximum drawdown. The strategies include (i) the long and short legs of the strategy based on ChatGPT 3.5, (ii) the long-short strategy based on ChatGPT 3.5, (iii) the long-short strategy based on ChatGPT 4, (iv) equal-weight and value-weight market portfolios, and (v) an equal-weight portfolio in all stocks with news the day before (regardless of news direction).

	Long (L)	Short (S)	L-S ChatGPT	L-S GPT-4	Market EW	Market VW	All News EW
Sharpe Ratio	1.72	1.86	3.09	3.80	-0.99	-0.39	-0.98
Daily Mean (%)	0.25	0.38	0.63	0.44	-0.10	-0.04	-0.11
Daily Std. Dev. (%)	2.32	3.26	3.25	1.84	1.55	1.49	1.83
Max Drawdown (%)	-16.94	-34.39	-22.79	-10.40	-36.12	-26.68	-38.70

Table 2: Regression of Next Day Returns on Prediction Scores from More Advanced LLMs

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 next day's return in percentage points, a_i and b_t are firm and time fixed effects, respectively. $x_{i,t}$ corresponds to the vector containing prediction scores from different models. The main regressors include scores from three advanced LLMs: (i) ChatGPT 3.5, (ii) ChatGPT 4, and (iii) BART Large. We include the event sentiment score from the data vendor for comparison purposes. We provide an overview of the different LLMs in Appendix A of the paper. 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 sample consists of all U.S. common stocks with at least one news headline covering the firm.

	(1)	(2)	(3)	(4)	(5)	(6)
GPT-3.5-score	0.259***	0.243***				
	(5.259)	(4.980)				
event-sentiment-score		0.058		0.038	0.118*	
		(1.122)		(0.683)	(2.272)	
GPT-4-score			0.176***	0.167***		
			(5.382)	(4.768)		
bart-large-score						0.142***
						(4.653)
Num.Obs.	60 755	60 755	60 755	60 755	60 755	60 176
R2	0.184	0.184	0.184	0.184	0.184	0.185
R2 Adj.	0.121	0.121	0.121	0.121	0.121	0.121
R2 Within	0.001	0.001	0.001	0.001	0.000	0.000
R2 Within Adj.	0.001	0.001	0.001	0.001	0.000	0.000
AIC	370 534.7	370 534.9	370 534.8	370 536.1	370 560.5	367 175.7
BIC	409 811.3	409 820.5	409 811.4	409 821.7	409 837.2	406 374.6
RMSE	4.75	4.75	4.75	4.75	4.75	4.76
Std.Errors	by: date & permno	by: date & permno	by: date & permno	by: date & permno	by: date & permno	by: date & permno
FE: date	X	X	X	X	X	X
FE: permno	X	X	X	X	X	X

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

Table 3: Regression of Next Day Returns on Prediction Scores from More Basic LLMs

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 next day's return in percentage points, a_i and b_t are firm and time fixed effects, respectively. $x_{i,t}$ corresponds to the vector containing prediction scores from different models. The main regressors include scores from six more basic LLMs: (i) DistilBart-MNLI-12-1, (ii) GPT-2 Large, (iii) GPT-2, (iv) GPT-1, (v) BERT, and (vi) BERT Large. We provide an overview of the different LLMs in Appendix A of the paper. 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 sample consists of all U.S. common stocks with at least one news headline covering the firm.

	(1)	(2)	(3)	(4)	(5)	(6)
distilbart-mnli-12-1-score	0.150***					
	(4.919)					
GPT-2-large-score		0.035				
		(1.051)				
GPT-2-score			0.001			
			(0.025)			
GPT-1-score				0.034		
				(1.304)		
bert-score					−0.226	
					(−3.703)	
bert-large-score						0.001
						(0.020)
Num.Obs.	60 755	60 176	60 176	60 755	60 176	60 176
R2	0.184	0.185	0.185	0.184	0.185	0.185
R2 Adj.	0.121	0.121	0.121	0.121	0.121	0.121
R2 Within	0.000	0.000	0.000	0.000	0.000	0.000
R2 Within Adj.	0.000	0.000	0.000	0.000	0.000	0.000
AIC	370 547.3	367 194.8	367 195.9	370 566.9	367 180.1	367 195.9
BIC	409 823.9	406 393.7	406 394.8	409 843.5	406 379.0	406 394.8
RMSE	4.75	4.76	4.76	4.75	4.76	4.76
Std.Errors	by: date & permno	by: date & permno	by: date & permno	by: date & permno	by: date & permno	by: date & permno
FE: date	X	X	X	X	X	X
FE: permno	X	X	X	X	X	X

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

Table 4: Regression of Next Day Returns on Prediction Scores from More Advanced LLMs (Small Stocks)

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 next day's return in percentage points, a_i and b_t are firm and time fixed effects, respectively. $x_{i,t}$ corresponds to the vector containing prediction scores from different models. The main regressors include scores from three advanced LLMs: (i) ChatGPT 3.5, (ii) ChatGPT 4, and (iii) BART Large. We include the event sentiment score from the data vendor for comparison purposes. 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 sample used in this analysis consists of Small stocks, defined as those whose market capitalization is below the 10th percentile NYSE market capitalization distribution.

	(1)	(2)	(3)	(4)	(5)	(6)
GPT-3.5-score	0.653***	0.542***				
	(5.145)	(4.028)				
event-sentiment-score		0.277*		0.256+	0.435***	
		(2.117)		(1.876)	(3.567)	
GPT-4-score			0.501***	0.419***		
			(4.830)	(3.645)		
bart-large-score						0.165
						(1.504)
Num.Obs.	14 343	14 343	14 343	14 343	14 343	14 238
R2	0.201	0.201	0.201	0.201	0.200	0.201
R2 Adj.	0.086	0.086	0.086	0.086	0.085	0.086
R2 Within	0.002	0.002	0.002	0.002	0.001	0.000
R2 Within Adj.	0.002	0.002	0.002	0.002	0.001	0.000
AIC	98 043.0	98 039.8	98 041.0	98 038.7	98 052.1	97 320.5
BIC	111 731.4	111 735.8	111 729.4	111 734.7	111 740.5	110 957.8
RMSE	6.51	6.51	6.51	6.51	6.51	6.50
Std.Errors	by: date & permno	by: date & permno	by: date & permno	by: date & permno	by: date & permno	by: date & permno
FE: date	X	X	X	X	X	X
FE: permno	X	X	X	X	X	X

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

Table 5: Regression of Next Day Returns on Prediction Scores from More Basic LLMs (Small Stocks)

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 next day's return in percentage points, a_i and b_t are firm and time fixed effects, respectively. $x_{i,t}$ corresponds to the vector containing prediction scores from different models. The main regressors include scores from six more basic LLMs: (i) DistilBart-MNLI-12-1, (ii) GPT-2 Large, (iii) GPT-2, (iv) GPT-1, (v) BERT, and (vi) BERT Large. 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 sample used in this analysis consists of Small stocks, defined as those whose market capitalization is below the 10th percentile NYSE market capitalization distribution.

	(1)	(2)	(3)	(4)	(5)	(6)
distilbart-mnli-12-1-score	0.207+					
	(1.895)					
GPT-2-large-score		0.019				
		(0.216)				
GPT-2-score			0.064			
			(0.765)			
GPT-1-score				0.008		
				(0.098)		
bert-score					−0.492**	
					(−2.598)	
bert-large-score						0.018
						(0.096)
Num.Obs.	14 343	14 238	14 238	14 343	14 238	14 238
R2	0.200	0.201	0.201	0.200	0.202	0.201
R2 Adj.	0.084	0.085	0.085	0.084	0.086	0.085
R2 Within	0.000	0.000	0.000	0.000	0.001	0.000
R2 Within Adj.	0.000	0.000	0.000	0.000	0.001	0.000
AIC	98 063.3	97 322.6	97 322.0	98 066.4	97 314.2	97 322.6
BIC	111 751.7	110 959.9	110 959.3	111 754.8	110 951.5	110 959.9
RMSE	6.51	6.50	6.50	6.51	6.50	6.50
Std.Errors	by: date & permno	by: date & permno	by: date & permno	by: date & permno	by: date & permno	by: date & permno
FE: date	X	X	X	X	X	X
FE: permno	X	X	X	X	X	X

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

51

Table 6: Regression of Next Day Returns on Prediction Scores from More Advanced LLMs (Non-Small Stocks)

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 next day's return in percentage points, a_i and b_t are firm and time fixed effects, respectively. $x_{i,t}$ corresponds to the vector containing prediction scores from different models. The main regressors include scores from three advanced LLMs: (i) ChatGPT 3.5, (ii) ChatGPT 4, and (iii) BART Large. We include the event sentiment score from the data vendor for comparison purposes. 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 sample consists of non-small stocks, defined as those whose market cap is above the 10th percentile NYSE market capitalization distribution.

	(1)	(2)	(3)	(4)	(5)	(6)
GPT-3.5-score	0.148**	0.158**				
	(3.084)	(3.280)				
event-sentiment-score		−0.041		−0.060	−0.005	
		(−0.830)		(−1.163)	(−0.112)	
GPT-4-score			0.097**	0.111***		
			(3.252)	(3.491)		
bart-large-score						0.144***
						(4.695)
Num.Obs.	46 402	46 402	46 402	46 402	46 402	45 928
R2	0.218	0.218	0.218	0.218	0.218	0.219
R2 Adj.	0.159	0.159	0.159	0.159	0.158	0.159
R2 Within	0.000	0.000	0.000	0.000	0.000	0.001
R2 Within Adj.	0.000	0.000	0.000	0.000	0.000	0.001
AIC	265 328.7	265 329.8	265 329.2	265 329.3	265 341.2	262 823.9
BIC	294 082.6	294 092.4	294 083.1	294 091.9	294 095.1	291 517.9
RMSE	3.93	3.93	3.93	3.93	3.93	3.94
Std.Errors	by: date & permno	by: date & permno	by: date & permno	by: date & permno	by: date & permno	by: date & permno
FE: date	X	X	X	X	X	X
FE: permno	X	X	X	X	X	X

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

Table 7: Regression of Next Day Returns on Prediction Scores from More Basic LLMs (Non-Small Stocks)

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 next day's return in percentage points, a_i and b_t are firm and time fixed effects, respectively. $x_{i,t}$ corresponds to the vector containing prediction scores from different models. The main regressors include scores from six more basic LLMs: (i) DistilBart-MNLI-12-1, (ii) GPT-2 Large, (iii) GPT-2, (iv) GPT-1, (v) BERT, and (vi) BERT Large. 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 sample consists of non-small stocks, defined as those whose market cap is above the 10th percentile NYSE market capitalization distribution.

	(1)	(2)	(3)	(4)	(5)	(6)
distilbart-mnli-12-1-score	0.146***					
	(4.894)					
GPT-2-large-score		0.030				
		(0.947)				
GPT-2-score			−0.014			
			(−0.539)			
GPT-1-score				0.056*		
				(2.332)		
bert-score					−0.165**	
					(−2.795)	
bert-large-score						−0.011
						(−0.200)
Num.Obs.	46 402	45 928	45 928	46 402	45 928	45 928
R2	0.218	0.219	0.219	0.218	0.219	0.219
R2 Adj.	0.159	0.158	0.158	0.158	0.159	0.158
R2 Within	0.001	0.000	0.000	0.000	0.000	0.000
R2 Within Adj.	0.001	0.000	0.000	0.000	0.000	0.000
AIC	265 316.8	262 848.3	262 848.9	265 336.7	262 839.8	262 849.1
BIC	294 070.7	291 542.3	291 542.8	294 090.6	291 533.7	291 543.1
RMSE	3.93	3.94	3.94	3.93	3.94	3.94
Std.Errors	by: date & permno	by: date & permno	by: date & permno	by: date & permno	by: date & permno	by: date & permno
FE: date	X	X	X	X	X	X
FE: permno	X	X	X	X	X	X

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

Table 8: Average Next Day's Return by Prediction Score

This table reports several statistics for the portfolios implied by different models. Columns Neg, Pos, and LS shows the daily average returns in percentage points (0.1 corresponds to 0.1%) for the long, neutral, short, and long-short portfolio. Column t LS shows the t-statistic of the daily returns for the long-short portfolio. α_M shows the daily alpha with respect to the CAPM model, t α_M is the t-statistic, and $R2_M$ is the R-sq. from the CAPM model in percentage points. 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 A of the paper. Panel A reports results for all U.S. common stocks with at least one news headline covering the firm, Panel B analyzes the sample of small stocks (below the 10th percentile NYSE market capitalization), and Panel C analyzes the remaining non-small stocks.

Panel A. Full Sample: All Stocks

Model	Pos	Neut	Neg	LS	t LS	α_M	t α_M	$R2_M$	α_{FF5}	t α_{FF5}	R^2_{FF5}
Gpt-4	0.09	-0.18	-0.35	0.44	4.24	0.45	4.31	1.14	0.41	4.01	5.20
Gpt-3.5	0.25	-0.21	-0.38	0.63	3.44	0.63	3.41	0.47	0.60	3.28	4.15
Gpt-1	-0.10	0.01	-0.19	0.09	0.69	0.09	0.71	0.24	0.09	0.67	0.41
Gpt-2	-0.03	-0.20	0.20	-0.23	-1.38	-0.23	-1.37	0.03	-0.24	-1.39	1.83
Gpt-2-Large	-0.06	-0.02	-0.16	0.10	0.92	0.10	0.95	0.40	0.10	0.94	0.64
Bart-Large	-0.01	0.08	-0.16	0.15	1.40	0.15	1.41	0.04	0.13	1.25	1.91
Distilbart-Mnli-12-1	-0.04	0.12	-0.28	0.24	2.12	0.24	2.13	0.08	0.22	1.91	3.77
Bert	-0.23	-0	-0.08	-0.14	-1.16	-0.12	-1.05	12.60	-0.09	-0.79	17.99
Bert-Large	-0.06	-0.06	-0.11	0.04	0.23	0.05	0.25	0.36	0.04	0.19	4.24
Event-Sentiment	-0.04	-0.11	-0.32	0.29	1.94	0.28	1.90	0.47	0.25	1.70	3.04
Finbert	-0.09	0.03	-0.21	0.11	0.86	0.12	0.87	0.05	0.13	0.96	2.21

Panel B. Small Stocks

Model	Pos	Neut	Neg	LS	t LS	α_M	t α_M	$R2_M$	α_{FF5}	t α_{FF5}	R^2_{FF5}
Gpt-4	0.02	-0.82	-0.88	0.90	3.13	0.90	3.12	0.88	0.88	3.07	3.88
Gpt-3.5	0.07	-0.51	-0.97	1.04	2.72	1.03	2.72	2.84	1.05	2.78	5.60
Gpt-1	-0.40	-0.31	-0.51	0.10	0.35	0.10	0.34	0.10	0.13	0.41	0.70
Gpt-2	-0.24	-0.81	-0.14	-0.10	-0.36	-0.10	-0.36	0	-0.11	-0.42	1.20
Gpt-2-Large	-0.23	-0.30	-0.39	0.16	0.55	0.16	0.55	0	0.20	0.68	0.65
Bart-Large	-0.19	-1.98	-0.48	0.29	1.09	0.29	1.10	0.02	0.28	1.06	1.04
Distilbart-Mnli-12-1	-0.25	-2.69	-0.46	0.20	0.73	0.20	0.73	0.11	0.21	0.74	1.26
Bert	-0.88	0.06	-0.08	-0.81	-4.09	-0.81	-4.27	6.57	-0.80	-4.17	8.32
Bert-Large	-0.15	-0.22	-0.54	0.40	1.30	0.39	1.29	1.17	0.42	1.38	3.58
Event-Sentiment	-0.26	-0.23	-0.79	0.53	2	0.53	2	0.12	0.57	2.11	0.64
Finbert	-0.47	-0.17	-0.43	-0.04	-0.12	-0.05	-0.14	2.06	-0.00	-0.01	3.20

Panel C. Non-Small Stocks

Model	Pos	Neut	Neg	LS	t LS	α_M	t α_M	$R2_M$	α_{FF5}	t α_{FF5}	R^2_{FF5}
Gpt-4	0.12	0.10	-0.20	0.32	3.32	0.33	3.37	0.71	0.29	2.99	4.07
Gpt-3.5	0.29	-0.06	-0.13	0.42	2.51	0.41	2.46	1.18	0.36	2.19	4.57
Gpt-1	0.02	0.06	-0.01	0.03	0.36	0.03	0.39	0.31	0.04	0.48	0.50
Gpt-2	0.06	-0	0.14	-0.08	-1.12	-0.07	-1.07	0.98	-0.07	-1.07	2.31
Gpt-2-Large	-0.02	0.10	-0.02	0	0.05	0.01	0.08	0.41	0.01	0.09	0.85
Bart-Large	0.10	0.32	-0.12	0.21	2.13	0.22	2.16	0.33	0.20	2.01	1.88
Distilbart-Mnli-12-1	0.09	0.40	-0.24	0.33	3.13	0.33	3.18	0.70	0.31	2.93	3.89
Bert	-0	0.01	-0.02	0.01	0.10	0.04	0.37	18.39	0.07	0.74	24.87
Bert-Large	0.02	0.12	0.18	-0.17	-1.06	-0.16	-1.01	1.02	-0.16	-1.03	6.41
Event-Sentiment	0.03	-0.03	0.04	-0.01	-0.05	-0.02	-0.10	1.01	-0.05	-0.33	3.83
Finbert	0.02	0.11	-0.09	0.10	0.89	0.10	0.88	0.00	0.11	0.93	2.57

Table 9: Sharpe Ratio and Number of Stocks in Each Leg by Model

This table reports the annualized Sharpe ratio of the long-short portfolio implied by different models. The table also reports the 25th percentile, mean, median, and 75th percentile of the number of stocks in the long (N_+) and in the short (N_-) legs. The models include ChatGPT 3.5, ChatGPT 4, Distilbart, Ravenpack, Bart-Large, Gpt-2-Large, Gpt-1, Bert-Large, Bert, Gpt-2. We provide an overview of the different LLMs in Appendix A of the paper. Panel A reports results for all U.S. common stocks with at least one news headline covering the firm, Panel B analyzes the sample of small stocks (below the 10th percentile NYSE market capitalization), and Panel C analyzes the remaining non-small stocks.

Panel A. Full Sample: All Stocks

Model	Sharpe	N_+25th	N_+mean	N_+median	N_+75th	N_-25th	N_-mean	N_-median	N_-75th
Gpt-4	3.8	59	98.52	107	140	13	23.30	23	33
Gpt-3.5	3.09	28	50.86	51	72	1	3.84	4	6
Distilbart-Mnli-12-1	1.9	104	167.87	160	234	11	22.06	21	32
Event-Sentiment	1.74	21	44.44	48	61	2	8.82	5	9
Bart-Large	1.26	21.50	148.04	150	210.50	4	23.71	24	34.25
Gpt-2-Large	0.82	8	59.84	61	83.50	2	12.37	12.50	18
Finbert	0.77	14.00	26.98	27.00	38.00	3.00	8.16	8.00	12.00
Gpt-1	0.62	89	138.23	136	187	14	26.03	25	36
Bert-Large	0.2	21.25	156.32	162.50	223.50	0	2.77	2	4
Bert	<0	7.75	36.17	43	53.25	0	0.27	0	0
Gpt-2	<0	15.75	99.55	104.50	144	3	24.25	25	35

Panel B. Small Stocks

Model	Sharpe	N_+25th	N_+mean	N_+median	N_+75th	N_-25th	N_-mean	N_-median	N_-75th
Gpt-4	2.98	19	25.52	26	32	2	4.24	4	6
Gpt-3.5	2.59	10	14.07	14	18	0	0.72	0	1
Event-Sentiment	1.91	8	12.63	13	16	0	3.63	1	3
Bert-Large	1.24	28	43.74	39	54	0	0.52	0	1
Bart-Large	1.04	26.25	44.65	37	56.75	2	4.31	4	6
Distilbart-Mnli-12-1	0.69	29	47.52	38	59	2	3.92	4	5
Gpt-2-Large	0.53	9	16.89	15	22	1	3.60	3	5
Gpt-1	0.33	23	37.32	32	48	4	7.24	6	9
Finbert	<0	3.00	6.34	6.00	9.00	0.00	1.51	1.00	2.00
Gpt-2	<0	16	28.53	25	36.75	4	6.92	6	9
Bert	<0	7	9.35	10	12.75	0	0.02	0	0

Panel C. Non-Small Stocks

Model	Sharpe	N_+25th	N_+mean	N_+median	N_+75th	N_-25th	N_-mean	N_-median	N_-75th
Gpt-4	2.99	49	76.40	80	108	9	19.66	19	28.50
Distilbart-Mnli-12-1	2.82	76	126.60	120	173	8.50	18.70	17	27
Gpt-3.5	2.26	20.50	38.63	37	54	1	3.22	3	5
Bart-Large	1.92	31	111.47	106.50	157	4.25	20.17	19	28
Finbert	0.8	11.50	21.50	21.00	30.00	2.00	6.86	6.00	10.00
Gpt-1	0.33	64	105.85	106	142	10	19.74	19	27.50
Bert	0.09	8.25	28.34	33	41	0	0.26	0	0
Gpt-2-Large	0.04	12.50	45.99	47.50	64	2	9.40	9	14
Event-Sentiment	<0	15	33.46	32	44	1.50	5.65	4	7
Bert-Large	<0	36	120.49	124	171.25	0	2.34	2	4
Gpt-2	<0	19.50	76	77	107.75	6.25	18.58	18	28

Table 10: Sharpe Ratios by News Complexity

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. The second 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.

Model	All	Low Complexity	High Complexity
Gpt-4	3.8	2.62	2.61
Gpt-3.5	3.09	1.53	1.66
Distilbart-Mnli-12-1	1.9	1.69	1.27
Event-Sentiment	1.74	1.51	0.48
Bart-Large	1.26	1.06	0.98
Gpt-2-Large	0.82	1.16	0.04
Finbert	0.77	0.97	0.31
Gpt-1	0.62	<0	0.35
Bert-Large	0.2	0.84	<0
Bert	<0	<0	<0
Gpt-2	<0	<0	<0

Table 11: Sharpe Ratios by 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 news articles, and the third column reports results using only press releases.

Model	All	News Articles	Press Releases
Gpt-4	3.8	4.74	2.84
Gpt-3.5	3.09	2.4	1.45
Distilbart-Mnli-12-1	1.9	4.8	0.73
Event-Sentiment	1.74	0.5	0.69
Bart-Large	1.26	3.74	1.37
Gpt-2-Large	0.82	0.83	1.34
Finbert	0.77	2.85	<0
Gpt-1	0.62	0.1	0.06
Bert-Large	0.2	0.82	0.02
Bert	<0	<0	<0
Gpt-2	<0	0.93	<0

Table 12: ChatGPT-4 Positive Recommendations Interpretability

This table reports on Panel A the most relevant words for making good predictions and their context. The coefficient is the slope from a regularized logistic regression model. The frequencies are normalized to the 0-1 range via TF-IDF, so the magnitude of the coefficient can be interpreted as feature importance. Larger coefficients are more relevant for accurate predictions.

Panel A: Positive Influence

Influential Word	Coefficient	Top Accompanying Words
purchase	0.61	future, shows, significant, number, demonstrates
guidance	0.50	indicate, revenue, stability, earnings, likely
share	0.39	earnings, market, indicate, typically, lead
dividends	0.37	generally, seen, sign, generating, profits
higher	0.35	lead, typically, attracts, indicate, sales
returns	0.31	shareholder, stability, attract, indicate, value
generating	0.28	profits, sign, generally, seen, sharing
number	0.26	significant, future, acquisition, shows, purchase
sharing	0.26	generating, profits, sign, generally, seen
insider	0.23	future, positively, significant, number, indicate

Panel B: Negative Influence

Influential Word	Coefficient	Top Accompanying Words
development	-0.54	progress, positively, new, investor, lead
profits	-0.45	generating, sign, generally, seen, sharing
stability	-0.28	sign, generally, seen, indicating, commitment
profitability	-0.27	announcement, shareholder, typically, quarterly...
sales	-0.24	indicate, likely, higher, boost, positively
commitment	-0.21	sign, generally, seen, shareholder, stability
declaring	-0.20	sign, seen, generally, quarterly, indicating
demand	-0.19	lead, increase, partnership, positively, likely
partnership	-0.18	likely, lead, revenue, boost, increase
lead	-0.16	revenue, typically, partnership, higher, collab...

Table 13: ChatGPT-4 Negative Recommendations Interpretability

This table reports on Panel A the most relevant words for making good predictions and their context. The coefficient is the slope from a regularized logistic regression model. The frequencies are normalized to the 0-1 range via TF-IDF, so the magnitude of the coefficient can be interpreted as feature importance. Larger coefficients are more relevant for accurate predictions.

Influential Word	Coefficient	Top Accompanying Words
significant	0.92	indicate, lack, selling, number, chairman
indicate	0.78	lack, significant, selling, number, future
risk	0.64	downgrade, credit, investor, higher, outlook
headline	0.48	suggests, likely, earnings, issues, sales
impacted	0.46	likely, earnings, revenue, reduced, drop
director	0.45	indicate, lack, number, sale, future
issues	0.43	headline, sales, impacting, revenue, reduced
number	0.39	lack, indicate, significant, selling, future
fraud	0.36	securities, reputational, investor, loss, headline
reputational	0.33	securities, fraud, losses, headline, lead

Influential Word	Coefficient	Top Accompanying Words
prospects	-0.90	lack, significant
credit	-0.63	downgrade, outlook, future, risk, investor
chairman	-0.62	indicate, lack, selling, significant, number
lack	-0.52	indicate, selling, number, significant, future
outlook	-0.47	downgrade, future, credit, investor, lowered
sale	-0.47	lack, indicate, number, future, director
revenue	-0.44	lower, likely, decreased, expectations, profit
earnings	-0.40	likely, impacted, lower, sales, decline
losses	-0.38	reputational, decreased, securities, impacts, lead
sales	-0.32	lower, decreased, indicate, decline, earnings

Online Appendix

Appendix A: Tables

Table 14: 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 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	5.26	−64.97	−2.04	−0.02	1.89	237.11	60755
Headline Length	76.36	28.65	21	56	70	90	395	60755
ChatGPT Response Length	153.31	38.04	0	124	151	179	303	60755
GPT Score	0.24	0.47	−1	0	0	1	1	60755
Event Sentiment Score	0.18	0.49	−1	0	0	0	1	60755

Panel B. Correlations

	Daily Return (%)	Headline Length	GPT Resp. Length	GPT Score	Event Sent. Score
Daily Return (%)	1				
Headline Length	−0.002	1			
ChatGPT Response Length	−0.001	0.261	1		
GPT Score	0.018	0.081	0.441	1	
Event Sentiment Score	0.005	−0.071	0.091	0.279	1

Appendix A: 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 GPT3 (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: 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. Event-Sentiment Estimated Number of Parameters: NA, Release Date: NA, Website: https://www.ravenpack.com/

RavenPack Event Sentiment Score

Model 11. FinBERT Estimated Number of Parameters: 774M, 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 B: 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.