

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 20, 2023

Abstract

We examine the potential of ChatGPT, and other large language models, in predicting stock market returns using sentiment analysis of news headlines. We use ChatGPT to indicate whether a given headline is good, bad, or irrelevant news for firms' stock prices. We then compute a numerical score and document a positive correlation between these "ChatGPT scores" and subsequent daily stock market returns. Further, ChatGPT outperforms traditional sentiment analysis methods. We find that more basic models such as GPT-1, GPT-2, and BERT cannot accurately forecast returns, indicating return predictability is an emerging capacity of complex models. Our results suggest that incorporating advanced language models into the investment decision-making process can yield more accurate predictions and enhance the performance of quantitative trading strategies.

*We are grateful for the comments and feedback from Andrew Chen, Carter Davis, Andy Naranjo, Nikolai Roussanov, and '@jugglingnumbers.' Emails: Alejandro Lopez-Lira (corresponding author): alejandro.lopez-lira@warrington.ufl.edu, and Yuehua Tang: yuehua.tang@warrington.ufl.edu.

The application of large language models (LLMs) such as ChatGPT in various domains has gained significant traction in recent months, with numerous studies exploring their potential in diverse areas. In financial economics, however, using LLMs remains relatively uncharted territory, especially concerning their ability to predict stock market returns. On the one hand, as these models are not explicitly trained for this purpose, one may expect that they offer little value in predicting stock market movements. On the other hand, to the extent that these models are more capable of understanding natural language, one could argue that they could be a valuable tool for processing textual information to predict stock returns. Thus, the performance of LLMs in predicting financial market movements is an open question.

To the best of our knowledge, this paper is among the first to address this critical question by evaluating the capabilities of ChatGPT in forecasting stock market returns. Through a novel approach that leverages the model's sentiment analysis capabilities, we assess the performance of ChatGPT using news headlines data and compare it to existing sentiment analysis methods provided by leading vendors.

Our findings have important implications for the employment landscape in the financial industry. The results could potentially lead to a shift in the methods used for market prediction and investment decision-making. By demonstrating the value of ChatGPT in financial economics, we aim to contribute to the understanding of LLMs' applications in this field and inspire further research on integrating artificial intelligence and natural language processing in financial markets. In addition to the implications for employment in the financial industry, our study offers several other significant contributions.

Firstly, our research can aid regulators and policymakers in understanding the potential benefits and risks associated with the increasing adoption of LLMs in financial markets. As these models become more prevalent, their influence on market behavior, information dissemination, and price formation will become critical areas of concern. Our findings can inform discussions on regulatory frameworks that govern the use of AI in finance and contribute to

the development of best practices for integrating LLMs into market operations.

Secondly, our study can benefit asset managers and institutional investors by providing empirical evidence on the efficacy of LLMs in predicting stock market returns. This insight can help these professionals make more informed decisions about incorporating LLMs into their investment strategies, potentially leading to improved performance and reduced reliance on traditional, more labor-intensive analysis methods.

Lastly, our research contributes to the broader academic discourse on artificial intelligence applications in finance. By exploring the capabilities of ChatGPT in predicting stock market returns, we advance the understanding of LLMs' potential and limitations within the financial economics domain. This can inspire future research on developing more sophisticated LLMs tailored to the financial industry's needs, paving the way for more efficient and accurate financial decision-making.¹

Our study has far-reaching implications that extend beyond the immediate context of stock market predictions. By shedding light on the potential contributions of ChatGPT to financial economics, we hope to encourage continued exploration and innovation in AI-driven finance.

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. Contemporaneously, Xie et al. (2023) find ChatGPT is no better than simple methods such as linear regression when using numerical data in prediction tasks.

1. See for example Wu et al. (2023).

We attribute the difference in results to their focus on using historical numerical data to predict, while ChatGPT excels at textual tasks. Ko and Lee (2023) finds ChatGPT may be useful in selecting across asset classes. Furthermore, Yang and Menczer (2023) demonstrates that ChatGPT successfully identifies credible news outlets. Our study is among the first to study the potential of LLMs in financial markets, particularly the investment decision-making process.

We contribute to the recent strand of the literature that employs text analysis and machine learning to study a variety of finance research questions (e.g., Jegadeesh and Wu (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), 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)). Our paper makes a unique contribution to this literature as being the first to evaluate the text processing capabilities of recently developed LLMs such as ChatGPT in forecasting stock market movements.

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 of the literature uses the sentiment of firm news to predict future individual stock returns (e.g., Tetlock, Saar-Tsechansky, and Macskassy (2008), Tetlock (2011), Jiang, Li, and Wang (2021)). Different from prior studies, we focus on understanding whether LLMs add value by extracting additional information that predicts stock market reactions.

Finally, our paper also relates to the literature on employment exposures and vulnerability to AI-related technology. Recent works by Agrawal, Gans, and Goldfarb (2019), Webb (2019), Acemoglu et al. (2022), Acemoglu and Restrepo (2022), Babina et al. (2022), 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. With AI being on a constant rise since its inception, our study focuses on understanding an urgent but unanswered question – the capabilities of AI, and LLMs in particular, in the finance domain. We highlight the potential of LLMs in adding value to market participants in processing information to predict stock returns.

1 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 to date 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 used 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 a range of natural language processing tasks, including language translation, text summarization, question answering, and text completion.

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

ChatGPT has been trained to perform a wide range of language tasks such as transla-

tion, summarization, question answering, and even 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 that can converse with users in a natural and intuitive way. While ChatGPT is a powerful tool for language-based tasks, it is not trained specifically to predict stock returns or provide financial advice. Hence, we test its capabilities when predicting stock returns.

2 Data

We utilize two primary datasets for our analysis: the Center for Research in Security Prices (CRSP) daily returns and news headlines from a leading data vendor. 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 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, providing a robust foundation for our analysis. Our sample consists of all the firms 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 the data vendor. Following prior studies, we use common stocks with a share code of 10 or 11.

We utilize news headlines from a prominent news sentiment analysis data provider, corresponding to the same time frame as the CRSP daily returns. The dataset comprises news headlines from a variety of sources, such as major news agencies, financial news websites, and

social media platforms. These headlines undergo preprocessing and filtering to emphasize company-specific news, enabling us to evaluate the influence of sentiment scores generated from these headlines on individual stock returns. We closely adhere to the preprocessing methods outlined by Jiang, Li, and Wang (2021).

We employ the “relevance score” provided, which ranges from 0 to 100, as an indicator of 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, and we limit it to full articles and press releases. We 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.

Furthermore, we eliminate duplicate headlines for the same company on the same day and extremely similar headlines. We gauge headline similarity using the Optimal String Alignment metric (also known as 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 any 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.

3 Methods

3.1 Prompt

Prompts are critical in guiding ChatGPT’s responses to specific tasks and queries. A prompt is a short piece of 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, generating a series of possible responses, 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 specific tasks in different domains, such as finance, healthcare, or customer support.

We use the following prompt in our study.

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 asked 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 provide a concise explanation in one sentence to support its answer. 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 a temperature of 0 to maximize the reproducibility of the results.

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

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 negative. However, ChatGPT’s response is that it believes the news to be positive for Oracle. ChatGPT’s reasoning is that the fine could increase investor confidence in Oracle’s ability to protect its intellectual property, potentially leading to an increase in 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.

3.2 Empirical Design

We prompt ChatGPT to provide a recommendation for each headline and transform it into a “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 lag the scores by one day to assess the predictability of returns. We then run linear regressions of the next day’s returns on the ChatGPT score and compare it to the sentiment score provided by a news curating company. Note that we are conservative in the news availability. If the news is reported after the closing time of the exchange, we assume the news is available to trade in the opening the day after. Thus, all of our results are out-of-sample.

4 Results

Our analysis reveals that ChatGPT sentiment scores exhibit a statistically significant predictive power on daily stock market returns. By utilizing news headline data and the generated sentiment scores, we find a strong correlation between the ChatGPT evaluation and the subsequent daily returns of the stocks in our sample. This result highlights the potential of ChatGPT as a valuable tool for predicting stock market movements based on sentiment analysis.

To further investigate the robustness of our findings, we compare the performance of ChatGPT with traditional sentiment analysis methods provided by a leading data vendor. In our analysis, we control for the ChatGPT sentiment 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 other sentiment scores on daily stock market returns is reduced to zero. 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.

Table 3 presents the results of our regression analysis, examining the relationship between next-day stock returns and sentiment scores generated by ChatGPT and alternative sentiment analysis methods. This table reports the regression coefficients and the corresponding t-statistics in parentheses. Standard errors are clustered by date and firm (permno).

The models include firm and date fixed effects to control for unobserved time-invariant firm characteristics and common time-specific factors that could influence stock returns. Various model fit measures, such as R-squared, adjusted R-squared, AIC, and BIC, are reported to assess the models' overall explanatory power.

We further present results for small stocks, defined as those smaller than the 10th percentile of the market cap of the NYSE, and non-small stocks, defined as the rest. The predictability is highly concentrated in small stocks, suggesting limits to arbitrage may limit the implementation and profitability of this strategy.

5 Conclusion

In this study, we have investigated the potential of ChatGPT, a large language model, in predicting stock market returns using sentiment analysis of news headlines. Our findings indicate that ChatGPT outperforms traditional sentiment analysis methods from a leading vendor. 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 should focus on understanding the mechanisms through which LLMs derive their predictive power. By identifying the factors that contribute to the success of models like ChatGPT in predicting stock market returns, researchers can develop more targeted strategies for improving these models and maximizing their utility in finance.

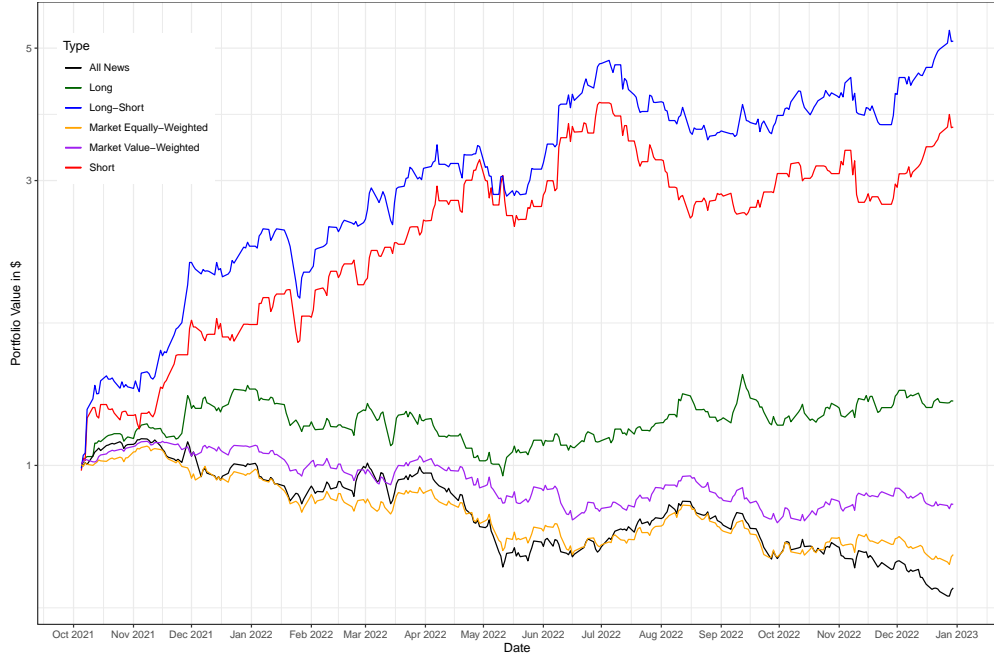
Additionally, as LLMs become more prevalent in the financial industry, it is essential to investigate their potential impact on market dynamics, including price formation, information dissemination, and market stability. Future research can explore the role of LLMs in shaping market behavior and their 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 and 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.

Figures

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



This figure presents the results of different trading strategies without considering transaction costs. We assume that if a piece of news is revealed before the market close, we buy (or short-sell) a position at the market close price. If a piece of news is announced after the market closes, we assume we buy (or short-sell) a position at the next opening price. 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. 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 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.

Tables

Table 1: Descriptive Statistics

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.

	Mean	SD	min	P25	Median	P75	Max	N
Daily Return (%)	-0.01	5.72	-64.97	-2.18	-0.04	1.96	237.11	39912
Headline Length	77.43	29.27	21	56	71	92	409	39912
ChatGPT Response Length	153	38.40	0	123	150	179	303	39912
GPT Score	0.24	0.47	-1	0	0	1	1	39912
Event Sentiment Score	0.18	0.50	-1	0	0	0	1	39912

Table 2: Correlations

This table reports the correlation between 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.

	Daily Return (%)	Headline Length	ChatGPT Response Length	GPT Score	Event Sentiment Score
Daily Return (%)	1
Headline Length	0.00	1	.	.	.
ChatGPT Response Length	0.00	0.26	1	.	.
GPT Score	0.02	0.08	0.44	1	.
Event Sentiment Score	0.00	-0.08	0.10	0.27	1

Table 3: Regression of Next Day Returns on the Prediction Score

This table reports the results of running regressions of the form $r_{i,t+1} = a_i + b_t + \gamma'x_t + \varepsilon_{i,t+1}$. Where $r_{i,t+1}$ is the next day's return in percentage points, a_i, b_t are firm and time fixed effects. x_t corresponds to the vector containing the ChatGPT or data vendor score. The corresponding t-statistics are in parentheses. Standard errors are clustered by date and firm. All models include firm and time fixed effects.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
GPT-score-a	0.278*** (4.477)	0.273*** (4.335)						
event-sentiment-score-a		0.022 (0.305)	0.085 (1.217)					
GPT-2-large-score-a				0.013 (0.308)				
GPT-2-score-a					-0.004 (-0.110)			
GPT-1-score-a						0.053 (1.509)		
BERT-large-score-a							0.066 (0.892)	
BERT-score-a								-0.287*** (-3.761)
Num.Obs.	39 912	39 912	39 912	39 912	39 912	39 912	39 912	39 912
R2	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
R2 Adj.	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118
R2 Within	0.001	0.001	0.000	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	0.000	0.000
AIC	250 298.1	250 300.0	250 317.5	250 319.7	250 319.8	250 317.8	250 319.1	250 305.0
BIC	276 296.3	276 306.7	276 315.7	276 317.9	276 317.9	276 315.9	276 317.2	276 303.2
RMSE	5.16	5.16	5.16	5.16	5.16	5.16	5.16	5.16
Std.Errors	by: date & permno	by: date & permno	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	X	X
FE: permno	X	X	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 the Prediction Score (Small Stocks)

This table reports the results of running regressions of the form $r_{i,t+1} = a_i + b_t + \gamma' x_t + \varepsilon_{i,t+1}$. Where $r_{i,t+1}$ is the next day's return in percentage points, a_i, b_t are firm and time fixed effects. x_t corresponds to the vector containing the ChatGPT or data vendor score. The corresponding t-statistics are in parentheses. Standard errors are clustered by date and firm. All models include firm and time fixed effects. Small stocks are defined as those whose market capitalization is less than the 10th percentile NYSE market capitalization.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
GPT-score-a	0.593*** (3.362)	0.514** (2.758)						
event-sentiment-score-a		0.202 (1.183)	0.346* (2.158)					
GPT-2-large-score-a				-0.046 (-0.404)				
GPT-2-score-a					0.046 (0.435)			
GPT-1-score-a						0.007 (0.063)		
BERT-large-score-a							0.097 (0.408)	
BERT-score-a								-0.570* (-2.370)
Num.Obs.	9941	9941	9941	9941	9941	9941	9941	9941
R2	0.210	0.210	0.209	0.209	0.209	0.209	0.209	0.209
R2 Adj.	0.085	0.085	0.084	0.084	0.084	0.084	0.084	0.084
R2 Within	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.001
R2 Within Adj.	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.001
AIC	69 419.8	69 420.1	69 425.9	69 431.2	69 431.2	69 431.4	69 431.2	69 424.3
BIC	79 196.2	79 203.7	79 202.3	79 207.6	79 207.6	79 207.8	79 207.6	79 200.7
RMSE	6.93	6.93	6.93	6.94	6.94	6.94	6.94	6.93
Std.Errors	by: date & permno	by: date & permno	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	X	X
FE: permno	X	X	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 Score (Non-Small Stocks)

This table reports the results of running regressions of the form $r_{i,t+1} = a_i + b_t + \gamma'x_t + \varepsilon_{i,t+1}$. Where $r_{i,t+1}$ is the next day's return in percentage points, a_i, b_t are firm and time fixed effects. x_t corresponds to the vector containing the ChatGPT or data vendor score. The corresponding t-statistics are in parentheses. Standard errors are clustered by date and firm. All models include firm and time fixed effects. Non-small stocks are defined as those whose market cap is greater than the 10th percentile NYSE market capitalization.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
GPT-score-a	0.174** (3.000)	0.187** (3.217)						
event-sentiment-score-a		-0.063 (-0.927)	-0.024 (-0.363)					
GPT-2-large-score-a				0.004 (0.103)				
GPT-2-score-a					-0.009 (-0.282)			
GPT-1-score-a						0.075* (2.390)		
BERT-large-score-a							0.035 (0.483)	
BERT-score-a								-0.229** (-3.048)
Num.Obs.	29 962	29 962	29 962	29 962	29 962	29 962	29 962	29 962
R2	0.219	0.219	0.219	0.219	0.219	0.219	0.219	0.219
R2 Adj.	0.154	0.154	0.154	0.154	0.154	0.154	0.154	0.154
R2 Within	0.000	0.000	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	0.000	0.000
AIC	176 382.5	176 383.3	176 391.8	176 392.0	176 392.0	176 387.7	176 391.8	176 381.6
BIC	195 407.1	195 416.2	195 416.4	195 416.6	195 416.6	195 412.3	195 416.4	195 406.2
RMSE	4.25	4.25	4.26	4.26	4.26	4.26	4.26	4.25
Std.Errors	by: date & permno	by: date & permno	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	X	X
FE: permno	X	X	X	X	X	X	X	X

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

Table 6: Selected Metrics

This table reports selected accuracy, precision, recall, specificity, and F1 score metrics. The table considers whether the firm’s stock market return is positive or negative. We only include observations where the model’s response is YES or NO (excluding UNKWON). The numbers are rounded to two decimals. Naive corresponds to predicting always the majority class.

Metric	GPT	sentiment	GPT-1	GPT-2	BERT-large	BERT	naive
Accuracy	0.51	0.51	0.50	0.50	0.50	0.50	0.50
Precision	0.51	0.51	0.50	0.50	0.51	0.50	0.50
Recall	0.93	0.92	0.86	0.86	0.98	1.00	1.00
Specificity	0.08	0.09	0.14	0.13	0.02	0.00	0.00
F1 Score	0.66	0.65	0.64	0.63	0.67	0.67	0.67

Table 7: Average Next Day’s Return by Prediction Score

This table reports the average daily returns in percentage points (0.1 corresponds to 0.1%) by the different model scores.

score	ChatGPT 3.5	GPT-1	GPT-2	BERT	Data Vendor
0	-0.05	-0.14	-0.12	0.05	-0.00
1	0.14	0.03	0.02	-0.23	-0.02
-1	-0.46	-0.10	0.10	-0.35	-0.11

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. ISSN: 0734306X. https://doi.org/10.1086/718327/SUPPL{_}FILE/20462DATA.ZIP.
- Acemoglu, Daron, and Pascual Restrepo. 2022. “Tasks, Automation, and the Rise in U.S. Wage Inequality.” *Econometrica* 90, no. 5 (September): 1973–2016. ISSN: 1468-0262. <https://doi.org/10.3982/ECTA19815>.
- 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. ISSN: 0895-3309. <https://doi.org/10.1257/JEP.33.2.31>.
- Babina, Tania, Anastassia Fedyk, Alex Xi He, and James Hodson. 2022. “Artificial Intelligence, Firm Growth, and Product Innovation.” *SSRN Electronic Journal* (May). <https://doi.org/10.2139/SSRN.3651052>.
- Baker, Scott R., Nicholas Bloom, and Steven J. Davis. 2016. “Measuring economic policy uncertainty.” *Quarterly Journal of Economics* 131, no. 4 (November): 1593–1636. ISSN: 15314650. <https://doi.org/10.1093/qje/qjw024>.
- 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. <https://doi.org/10.3386/w27843>.
- Bybee, Leland, Bryan T. Kelly, Asaf Manela, and Dacheng Xiu. 2019. “The Structure of Economic News.” *Working Paper* (January). ISSN: 1556-5068. <https://doi.org/10.2139/ssrn.3446225>.

- Bybee, Leland, Bryan T. Kelly, Asaf Manela, and Dacheng Xiu. 2021. “Business News and Business Cycles.” *SSRN Electronic Journal* (September). ISSN: 1556-5068. <https://doi.org/10.2139/SSRN.3446225>.
- 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. ISSN: 0304-405X. <https://doi.org/10.1016/J.JFINECO.2018.11.009>.
- 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. ISSN: 1380-6653. <https://doi.org/10.1007/S11142-013-9258-3/TABLES/11>.
- Cohen, Lauren, Christopher Malloy, and Quoc Nguyen. 2020. “Lazy Prices.” *Journal of Finance* 75 (3): 1371–1415. ISSN: 15406261. <https://doi.org/10.1111/jofi.12885>.
- Cowen, Tyler, and Alexander T. Tabarrok. 2023. “How to Learn and Teach Economics with Large Language Models, Including GPT.” *SSRN Electronic Journal* (March). ISSN: 1556-5068. <https://doi.org/10.2139/SSRN.4391863>.
- Freyberger, Joachim, Andreas Neuhierl, and Michael Weber. 2020. “Dissecting Characteristics Nonparametrically.” *The Review of Financial Studies* 33 (5): 2326–2377. ISSN: 0893-9454. <https://doi.org/10.1093/rfs/hhz123>.
- Garcia, Diego. 2013. “Sentiment during Recessions.” *The Journal of Finance* 68, no. 3 (June): 1267–1300. ISSN: 1540-6261. <https://doi.org/10.1111/JOFI.12027>.
- Gaulin, Maclean Peter. 2017. “Risk Fact or Fiction: The Information Content of Risk Factor Disclosures.”

- Gu, Shihao, Bryan Kelly, and Dacheng Xiu. 2020. “Empirical Asset Pricing via Machine Learning.” *The Review of Financial Studies* 33 (5): 2223–2273. ISSN: 0893-9454. <https://doi.org/10.1093/rfs/hhaa009>.
- Hansen, Anne Lundgaard, and Sophia Kazinnik. 2023. “Can ChatGPT Decipher FedSpeak?” *SSRN Electronic Journal* (March). ISSN: 1556-5068. <https://doi.org/10.2139/SSRN.4399406>.
- 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. ISSN: 0033-5533. <https://doi.org/10.1093/qje/qjx045>.
- Hoberg, Gerard, and Gordon Phillips. 2016. “Text-Based Network Industries and Endogenous Product Differentiation.” *Journal of Political Economy* 124 (5): 1423–1465. <https://doi.org/10.1086/688176>.
- Jegadeesh, Narasimhan, and Di Wu. 2013. “Word power: A new approach for content analysis.” *Journal of Financial Economics* 110 (3): 712–729. ISSN: 0304-405X. <https://doi.org/https://doi.org/10.1016/j.jfineco.2013.08.018>.
- 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. ISSN: 0304-405X. <https://doi.org/10.1016/J.JFINECO.2021.04.003>.
- 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*, <https://doi.org/http://dx.doi.org/10.2139/ssrn.3074808>.

- Ko, Hyungjin, and Jaewook Lee. 2023. “Can Chatgpt Improve Investment Decision? From a Portfolio Management Perspective.” *SSRN Electronic Journal*, <https://doi.org/10.2139/SSRN.4390529>.
- Korinek, Anton. 2023. “Language Models and Cognitive Automation for Economic Research.” (Cambridge, MA) (February). <https://doi.org/10.3386/W30957>.
- Lopez-Lira, Alejandro. 2019. “Risk Factors That Matter: Textual Analysis of Risk Disclosures for the Cross-Section of Returns.” *SSRN Electronic Journal* (September). ISSN: 1556-5068. <https://doi.org/10.2139/ssrn.3313663>.
- Manela, Asaf, and Alan Moreira. 2017. “News implied volatility and disaster concerns.” *Journal of Financial Economics* 123, no. 1 (January): 137–162. ISSN: 0304405X. <https://doi.org/10.1016/j.jfineco.2016.01.032>.
- Noy, Shakked, and Whitney Zhang. 2023. “Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence.” *SSRN Electronic Journal* (March). <https://doi.org/10.2139/SSRN.4375283>.
- 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. ISSN: 1540-6261. <https://doi.org/10.1111/J.1540-6261.2007.01232.X>.
- . 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. ISSN: 0893-9454. <https://doi.org/10.1093/RFS/HHQ141>.
- 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. ISSN: 15406261. <https://doi.org/10.1111/j.1540-6261.2008.01362.x>.

- 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. ISSN: 10495258.
- Webb, Michael. 2019. “The Impact of Artificial Intelligence on the Labor Market.” *SSRN Electronic Journal* (November). <https://doi.org/10.2139/SSRN.3482150>.
- 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” (April). <https://arxiv.org/abs/2304.00228v1>.