ChatGPT, Stock Market Predictability and Links to the Macroeconomy*

Jian Chen† Guohao Tang‡ Guofu Zhou§ Wu Zhu¶

First Draft: July 2023

Current Version: December 2023

*We are grateful to Ilias Filippou, Alejandro Lopez-Lira, Jun Pan, Yuehua Tang, Dacheng Xiu, Xintong Zhan, Dexin Zhou, and seminar participants at Dongbei University of Finance and Economics, Hunan University, New York University (Shanghai), Xiamen University, and Zhejiang University.

†School of Economics & Paula and Gregory Chow Institute for Studies in Economics, Xiamen University, China, 361005; e-mail: jchenl@xmu.edu.cn.

‡College of Finance and Statistics, Hunan University, China, 410006; e-mail: ghtang@hnu.edu.cn.

§**Corresponding author**: Olin School of Business, Washington University in St. Louis, St. Louis, Missouri, USA, 63130; e-mail: zhou@wustl.edu.

¶Department of Finance, School of Economics and Management, Tsinghua University, China, 100084; e-mail: zhuwu@sem.tsinghua.edu.cn

ChatGPT, Stock Market Predictability and Links to the Macroeconomy

Abstract

This paper examines whether ChatGPT can identify useful news content for the aggregate stock market and macroeconomy, using the news headlines and alerts on front pages of Wall Street Journal. We find that the information extracted by ChatGPT is highly related to macroeconomic conditions. Investors tend to underreact to the positive contents, especially during periods of economic downturns, high information uncertainty and high novelty of news, which leads to significant market predictability by ChatGPT. By contrast, the negative news is only associated with contemporaneous returns, and it cannot predict future market. Traditional methods of textual analysis, such as word lists or small large language models (LLMs) like BERT, can barely find any predictability in either positive news nor negative news. In short, ChatGPT appears the best of its kind and is capable of discerning economic-related news that drive the stock market.

JEL classifications: C22, C53, G11, G12, G17

Keywords: LLMs, ChatGPT, Textual Analysis, NLP, Return Predictability

1. Introduction

How information is incorporated into asset prices is a core issue in finance. In the age of big data, the amount of information being produced has grown rapidly, increasing the complexity of processing the information. In recent decades, with the development of natural language processing (NLP) technique, financial economics has begun to extract information about stock market from various text sources, like financial news press (see a comprehensive review of Loughran and McDonald, 2020). Nevertheless, unlike the regular structure of numerical data, the text data is far more difficult to work with. Recent textual analysis in finance and economics is the tip of the iceberg (Chen, Kelly, and Xiu, 2023). Thus, the disclosed textual information may not be fully incorporated by human investors and researchers with using conventional methods, e.g., bag-of-words or dictionary-based sentiment score. Comparing with these conventional methods, large language models (LLMs), which leverage pre-trained language models or word embeddings to enhance the performance across diverse NLP tasks significantly (Devlin, Chang, Lee, and Toutanova, 2018; Rogers, Kovaleva, and Rumshisky, 2021), may capture both the syntax and semantics of text substantially better.

In this paper, we investigate the comparative ability of LLMs, specifically ChatGPT, in capturing textual information vis-à-vis human interpretation using traditional methods. The primary inquiry revolves around whether ChatGPT surpass human capacity in information extraction from textual data. If confirmed, the implication suggests that information derived from LLMs may not be immediately integrated into stock prices by human investors, leading to a gradual market response to this information. To explore this issue, we download the news headlines and alerts on front page of *Wall Street Journal* from 1996 to 2022, and ask ChatGPT-3.5 to identify good and bad news by inputting a prompt: "*Forget all previous instructions. You are now a financial expert giving investment advice. I'll give you a news headline, and you need to answer whether this headline suggests the U.S. stock prices are GOING UP or GOING DOWN. Please choose only one option from GOING UP, GOING DOWN, UNKNOWN, and do not provide any additional responses.*" Next, we compute the monthly ratio of good news to total news and monthly ratio of bad news to total news. Using these two ratios of good and bad news, we examine their return predictability on the aggregate stock market.

Empirically, we find a high ratio of good news is positively correlated with contemporaneous market returns, and it significantly predicts subsequent high returns up to next six months for the sample period from January 1996 to December 2022. The R^2 of the regression of one-month ahead market excess return on good news ratio (NR^G) is 1.37%, with a slope of 0.53% which is statistically significant at the 5% level. With the increase of prediction horizon, the R^2 rises and reaches 8.52% over the annual horizon. The positive predictive power suggests an underrecation of investors to information of good media news. Additionally, the significantly positive relationship between current return and NR^G, indicating a gradual market response, rules out, in our context, the possible interpretations of information diffusion with a delay (Hong and Stein, 1999; Green, Huang, Wen, and Zhou, 2019; Agrawal, Hacamo, and Hu, 2021) and inattention of investors (e.g., DellaVigna and Pollet, 2009; Hirshleifer, Lim, and Teoh, 2009; Ben-Rephael, Da, and Israelsen, 2017; Anastassia, 2023).

In contrast to results for good news, we find no evidence of forecasting power for bad news ratio. As expected, it is correlated with contemporaneous returns negatively, suggesting that investors capture and react to the information of bad news. The process appears efficient, as there is no future implications. Overall, our results reveal that ChatGPT can capture more effectively textual information of good news than human does, leading to a return predictability of NR^G.

In a comparison with conventional methods of textual analysis, we find that the word lists proposed by Loughran and McDonald (2011) cannot predict the market in our new analysis. A plausible explanation for this discrepancy could be the diminished novelty of information identified through word lists to human investors, given the widespread adoption of this method since its introduction in 2011. Next, we check whether alternative LLMs can identify the textual information and deliver forecasting power for the aggregate stock market. We use the BERT and the RoBERTa developed by *google* to identify the good and bad news from the front-page news on *Wall Street Journals*. We find limited forecasting power of these two alternative LLMs. These observations are congruent with recent findings (Brown et al., 2020; Wei et al., 2022; Wei et al., 2022; Zhao et al., 2023), which suggest that "large" language models, endowed with hundreds of billions or more parameters manifest certain "emergent

abilities" not present in "smaller" models such as BERT. Additionally, in a comparison with common economic predictors and lagged market returns, we find that the predictability of NR^G remains strong. This finding may alleviate the concerns that the news content in *Wall Street Journal* might already be encapsulated within economic fundamentals or the return predictability comes from the continuation of stick prices.

Our findings are robust in quite a few dimensions. First, the predictability exists for alternative prompts. We use keywords of "*POSITIVE*" ("*OPTIMISTIC*" or "*GOOD*") and "*NEGATIVE*" ("*PESSIMISTIC*" or "*BAD*") instead of "*GOING UP*" and "*GOING DOWN*" to ask ChatGPT-3.5. The newly defined NR^G still predicts the market strongly. Second, we use ChatGPT-3.5 fine-tuning and ChatGPT-4, respectively, to identify good and bad news, and find that the return predictability remains significant. Third, the strong return predictability of NR^G also exists out-of-sample, which has become a critical assessment of predictability ever since Welch and Goyal (2008), because in-sample predictability can be unreliable due to parameter instability. Following the predictability literature, we evaluate Campbell and Thompson (2008)'s out-of-sample R^2_{OS} statistic, and find that NR^G still delivers statistically significant R^2_{OS} of 1.17% for the out-of-sample period from January 2006 to December 2022.

There is also significant economic value of the predictability of NR^G. Because of the significant positive R^2_{OS}, a mean-variance investor who allocates funds monthly between the market and risk-free assets can earn investment gains if the investor uses return forecasts based on NR^G rather than using the historical average return. Indeed, the annualized certainty equivalent return (CER) gain is 4.92% if the investor has a risk aversion degree of 3. This investment profit remains sizable after considering a proportional transaction cost of 50 basis points. The net-of-transaction-cost CER gains of NR^G is 3.55%. Moreover, the return forecasts of NR^G generate a large annualized Sharpe ratio of 0.51, while the market has a Sharpe ratio of only 0.30. Our results are robust to alternative risk aversion coefficients, such as one or five.

We explore further possible underlying economic mechanisms for the gradual market response to good news. First, we examine the information content estimated by ChatGPT and find that it is more likely to be related to macroeconomic condition. Specifically, a higher ratio of good news is indicative of an improving economic state, whereas a higher ratio of

bad news suggests deteriorating future economic conditions. This outcome evidences the capability of ChatGPT in capturing pertinent information about the macroeconomic state from news text, through which it influences the aggregate stock market. Second, we conjecture that human investors may not fully capture the textual information of good news during economic downturns. Veronesi (1999) shows that the market response to news can be dependent on the state of business cycle, and in particular, stock prices underreact to good news in bad times. Similarly, García (2013) find that the predictability of stock returns using news' content is concentrated in recessions. We then employ the Chicago Fed National Activity Index (CFNAI) as the proxy of economic condition, which assesses the overall economic activity and related inflationary pressure. We construct an indicator variable I_{High} that equals one if current CFNAI is above the past five-year sample mean, and zero otherwise; and let $I_{Low} = 1 - I_{High}$. Results of regression of future market returns on interaction terms between NR^G and I_{High} and between NR^G and I_{Low} show that the forecasting power of NR^G is more significant during periods of low economic activity, confirming our conjecture. Second, the literature suggests that investors underreact to news when information uncertainty is high (Zhang, 2006). Thus, the return predictability of NR^G might be stronger in high uncertainty of good news than in low uncertainty. We find indeed such evidence in regressing future returns on interaction terms between NR^G and dummy variables based on economic policy uncertainty (EPU) which is downloaded from Baker, Bloom, and Davis (2016). Third, we hypothesize that investors are likely to underreact to new information. Chan, Jegadeesh, and Lakonishok (1996) show that market response to recently released information is gradual. Daniel, Hirshleifer, and Subrahmanyam (1998) develop a model in which investors are overconfident with their private information and therefore underreact to public signals. Following the spirit of Tetlock (2011), we measure the novelty of information by using similarity of news stories to prior news stories relevant for economy. Again, we construct interaction terms between NR^G and dummy variables based on the similarity. We find consistent evidence with our conjecture that the return predictability of NR^G is stronger for low similarity of news than for high similarity.

Our paper contributes to the study of the revolutionary language and artificial intelligence (AI) platform, ChatGPT, which was released on November 30, 2022 and has soon

reached over 100 million users as of January 2023 and has growing influence throughout the world each day.[1] Due to its increasing impact on the whole economy and the AI community, it is of great interest and importance for understanding its implications in finance. Lopez-Lira and Tang (2023) is the first to study the predictive power of ChatGPT on stock returns, whereas Chen, Kelly, and Xiu (2023) use BERT to examine the cross-section of expected stock returns. In contrast to these studies which are about individual stock returns, we focus on the forecasting ability of ChatGPT on the aggregate stock market. Our methods, results, and the explanations are quite different from other studies as well.

Our paper also contributes to the literature on textual analysis. In the field of finance and accounting, textual analysis plays a pivotal role, scrutinizing text through the lenses of readability, similarity, and sentiment. Earlier studies are Antweiler and Frank (2004), Tetlock (2007), Tetlock, Saar-Tsechansky, and Macskassy (2008), Li (2008), and Tetlock (2011), among others. Since Loughran and McDonald (2011), their method of dictionary sentiment score has been widely used by the literature, e.g., García (2013), Jiang, Lee, Martin, and Zhou (2019), and Cohen, Malloy, and Nguyen (2020). Improving this method, the literature proposes unsupervised topic models (Bybee, Kelly, Manela, and Xiu, 2023; Bybee, Kelly, and Su, 2023) and supervised models (Manela and Moreira, 2017; García, Hu, and Rohrer, 2023). In contrast to these studies, we use large language models (LLMs), exemplified by ChatGPT, to extract the textual information. Comparing with the methods proposed by the prior studies, LLMs can capture both the syntax and semantics of text substantially better.

The remainder of the paper is organized as follows: Section **2** briefly introduces the large language models and their backgrounds. Section **3** describes the empirical methods and data. Section **4** provides our main empirical results. Section **5** explores the economic mechanism of predictability. Section **6** concludes the paper.

2. Large Language Models

This section briefly describes the background and the recent development of Natural Language Processing and Large Language Models (LLMs). The primary objective of NLP en-

[1] https://www.cnn.com/2023/11/30/tech/chatgpt-openai-revolution-one-year/index.html

compasses the development of robust and versatile representations for linguistic units such as words, sentences, paragraphs, and even larger text structures. These representations form the backbone for various downstream NLP applications, ranging from text classification and information extraction to entity recognition and question-answering systems. Recent advancements in this domain have been propelled by the advent of Large Language Models (LLMs), which leverage pre-trained language models or word embeddings to enhance the performance across diverse NLP tasks significantly (Devlin, Chang, Lee, and Toutanova, 2018; Rogers, Kovaleva, and Rumshisky, 2021). We use a suite of state-of-the-art LLMs, including ChatGPT-3.5, ChatGPT-4 (Radford et al., 2019), BERT (Devlin et al., 2018), and RoBERTa (Liu et al., 2019) to derive vector representations of news headlines. These models, with their distinct architectures and training methodologies, provide a comprehensive basis for analyzing and interpreting the semantic and syntactic nuances embedded in textual data.

Recent advancements in Large Language Models (LLMs) have introduced a paradigm shift in contrast with the paradigm of traditional word embedding, which assigns a static vector representation to each word in a predefined vocabulary irrespective of its contextual usage. These models employ attention mechanisms, a notable example being the Transformer architecture (Vaswani et al., 2017), to dynamically learn word representations based on the surrounding context in which a word appears (Peters et al., 2018). This contextualized approach enables a more nuanced understanding of language, as the meaning of words can vary significantly depending on their usage in different sentences (Peters et al., 2018; Radford et al., 2019).

Here, we briefly describe two classes of LLMs-BERT and -ChatGPT and their applications in our context. Both models are built on the transformer (Vaswani et al., 2017). Our empirical analysis make comparisons across models to examine how model size, robust adjustment, and fine-tuning in the context of financial documents affect the baseline results.

2.1. BERT

Bidirectional Encoder Representations from Transformers (BERT) is a transformers model pre-trained on a large corpus of text data in a self-supervised way (Devlin, Chang, Lee, and Toutanova, 2018). The basic architecture of the BERT is a multi-layer bidirectional

Transformer encoder based on the original implementation by Vaswani et al. (2017). Each transformer layer contains two sub-components: a multi-head self-attention mechanism and a fully connected feed-forward network. This design allows BERT to analyze and understand the context of a word based on all other words in a sentence, which is a significant departure from previous models that processed text unidirectionally. Using the attention mechanism, the embedding of each word depends on the context in which the word appears, in contrast to the traditional word embedding where each word is assigned as a fixed vector representation. Moreover, the Transformer architecture supplants the sequential processing typical of recurrent neural networks (RNNs) with parallel processing attention mechanisms. This feature enables the model to assess the relevance of each word in a sentence without being constrained by their positional relationships.

BERT's pre-training involves two distinct tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In MLM, the model randomly masks 15% of the words in each input sentence and then runs the entire masked sentence through the model and tries to predict the masked words by processing the entire sentence. This task is designed to enable the model to learn the word representation based on the left and right context. In NSP task, the model concatenates two masked sentences as inputs during pre-training. Sometimes they correspond to sentences that were next to each other in the original text, sometimes not. The model then has to predict if the two sentences were following each other or not. Through this process, the model learns an inner representation (embedding) of words, sentences, and paragraphs, and the embed-dings are then used to extract useful features for downstream analysis.

The initial BERT model includes two versions: BERT BASE and BERT LARGE. Both take the same transformer structure but with different size of parameters. The BERT BASE consists of 110M while BERT LARGE includes 340M parameters. RoBERTa, developed by Liu et al. (2019), emerges as an optimized iteration of BERT, enhancing its training methodology and dataset size to achieve improved performance across a range of NLP benchmarks.

2.2. *ChatGPT*

The Generative Pre-trained Transformer (GPT) series, developed by OpenAI, represents a significant milestone in NLP and artificial intelligence (AI). These models are a part of the broader family of Transformer-based models introduced by Vaswani et al. (2017) and Radford et al. (2019). Similar to the BERT, at its core, each GPT model is based on the Transformer architecture. Unlike traditional recurrent neural network (RNN) models that process input data sequentially, Transformers process all parts of the input data in parallel, significantly improving efficiency and scalability. This architecture allows GPT models to effectively capture the context dependencies of words and sentences. GPT models are characterized by their large-scale unsupervised pre-training. They are initially trained on vast amounts of text data, learning to predict the next word in a sentence. This pre-training phase enables the models to understand language patterns and structures deeply.

Compared to the BERT model, a striking feature of the ChatGPT is the significant scaling up in model parameters and training data. For example, ChatGPT-3 takes 175B parameters, GPT2 takes 1.5B parameters, but BERT only has 340M parameters. As the parameter scale of language models surpasses a defined threshold, a marked enhancement in performance becomes evident. Moreover, these advanced, large-scale models manifest distinctive capabilities, such as in-context learning, which remain absent in their smaller-scale counterparts. For example, ChatGPT-3 can solve few-shot tasks through in-context learning or sequential reasoning in complex tasks, while GPT-2 or BERT can not do that well. This delineation suggests that beyond quantitative improvements, qualitative transformations in model abilities emerge as a function of increased scale, which was documented as the *emergent abilities* (Brown et al., 2020; Wei et al., 2022; Zhao et al., 2023).

The concept of *emergent abilities* in LLMs is rigorously delineated as capabilities absent in smaller models, such as BERT or GPT-2, but manifesting in more extensive models with hundreds or even thousands of billions of parameters (Wei et al., 2022; Zhao et al., 2023). These sources enumerate several *emergent abilities*, including instruction tuning, in-context learning, and step-by-step reasoning. Instruction tuning is characterized by the LLMs' ability to adapt to new tasks guided solely by instructions, sans explicit examples, thereby

exhibiting enhanced generalization. Within the ChatGPT series, this ability, also known as prompt engineering, forms the baseline of our empirical studies. In-context learning refers to the LLMs' capability to produce the anticipated output under natural language instructions or through a few task demonstrations without necessitating further training or gradient updates. Step-by-step reasoning, a concept highlighted in Wei et al. (2022), describes the LLMs' proficiency in navigating multi-step reasoning challenges using a chain-of-thoughts (COT) approach.

In our empirical analysis, we initially apply instruction tuning to assess ChatGPT-3.5's capability in extracting pertinent information for stock market prediction. Subsequently, we employ a range of fine-tuning techniques and instructions to evaluate the robustness of our baseline findings.

3. Empirical Methods and Data

3.1. Prompt

Our baseline is to take advantage of the instruction ability of ChatGPT-3.5 to extract relevant information to predict the aggregate stock market in the future. Specifically, News headlines are directly inputted into the model via a carefully designed prompt. This prompt functions as a directive mechanism, channeling the model's focus towards generating stock market predictions. In the operational framework of GPT models, a prompt is not merely a query or instruction posed to the model; it encompasses a broader scope, often integrating contextual information, specific input data, or illustrative examples to guide the model's response more effectively (Brown et al., 2020). This approach is predicated on the hypothesis that ChatGPT-3.5's advanced language processing and pattern recognition capabilities can discern and extrapolate relevant financial information and indicators from the textual data presented in news headlines.

We input the News headlines and alerts from *Wall Street Journal* from 1996 to 2022 and ask ChatGPT-3.5 (our baseline) to identify the good and bad news for the stock market. The prompt is:

"*Forget all previous instructions. You are now a financial expert giving investment advice. I'll give you a news headline, and you need to answer whether this headline suggests the U.S. stock prices are GOING UP or GOING DOWN. Please choose only one option from GOING UP, GOING DOWN, UNKNOWN, and do not provide any additional responses.*"

As such, we count the number of good and bad news items each month and compute the percentage number (the number of good or bad news items divided by the total number of news items within a month). We use this percentage number to predict stock market returns.

Compared to ChatGPT-3.5, ChatGPT-4 significantly expands the model size and offers greater accuracy and reliability across various language tasks like prediction and text understanding. The OpenAI finds that ChatGPT-4 can deal more with complex, multifaceted scenarios, which could be the case in finance, economics, and stock market analysis. To make a comparison, we use the same prompt as in ChatGPT-3.5 to judge the direction of stock movement based on the news headlines.

3.2. Few Shots

In our baseline, we implemented a "zero-shot" prompting technique with ChatGPT-3.5, where the model was prompted to render judgments without access to any representative news headlines or their corresponding classifications. This method exemplifies zero-shot learning, in which large language models have shown remarkable adeptness. Nevertheless, it is observed that their efficacy often declines when applied to more intricate tasks under zero-shot settings (Kaplan et al., 2020; Brown et al., 2020; Touvron et al., 2023). One can use the "few-shot" prompting method to amplify the model's capability for in-context learning. This approach integrates specific examples within the prompt, facilitating improved model performance.

In our study, one challenge arises from the fact that over 80% of news headlines are typically categorized as *UNKNOWN*. In a balanced few-shot scenario, at least ten examples are required to guarantee the presence of the *GOING UP* and *GOING DOWN* categories. However, incorporating numerous examples in a "few-shot" setup could lead to prohibitive token consumption per prompt. To circumvent this issue, we opted for direct model fine-

tuning, tailoring the model parameters to better align with our specific research objectives.

3.3. Fine-Tuning

In our study, a primary challenge in employing few-shot prompts for more precise classification is the potential mismatch in example distribution between the prompt and the actual data. Notably, *UNKNOWN* classifications account for over 80% of news headlines. Implementing a few-shot prompt with balanced classes (comprising two instances each of *GOING UP*, *GOING DOWN*, and *UNKNOWN*) may inadvertently bias the model away from the "unknown" category.

As OpenAI has elucidated, 'fine-tuning surpasses few-shot learning by training on a more extensive set of examples than what is feasible within a prompt, thereby enabling the model to excel in a broader array of tasks.' Building on this insight, we explore whether fine-tuning ChatGPT-3.5 enhances model performance. To this end, we selected a random subset of 300 news headlines, manually and carefully classified each into *GOING UP*, *GOING DOWN*, and *UNKNOWN*, and used this subset for fine-tuning ChatGPT-3.5's parameters.

In contrast to GPT models, BERT-type models do not adhere to a question-answer or completion format. To assess the predictive capabilities of BERT-type models in stock market analysis, we directly fine-tuned the BERT and ROBERTA models using the same 300 manually labeled news headlines. We divided this dataset into training (80%) and validation (20%) segments. The selection of hyperparameters—including early stopping (epochs), dropout rate, and parameter shrinkage—was guided by validation accuracy.

3.4. Embedding and News Similarity

In addition to the sentiment classification of news headlines, contemporary Large Language Models (LLMs) facilitate the derivation of vector representations for each headline. It significantly differs from traditional vector representation methods that predominantly rely on word frequency counts. Modern LLMs, incorporating sophisticated attention mechanisms, enable the incorporation of contextual information within sentence representations, thereby enriching the semantic understanding of the text. In our study, we employ a suite of advanced

models - BERT (Devlin, Chang, Lee, and Toutanova, 2018), RoBERTa (Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer, and Stoyanov, 2019), ChatGPT-3.5, and ChatGPT-4 - for the extraction of vector representations from news headlines. These representations are subsequently utilized to compute the similarity of current news items relative to preceding ones, offering a nuanced approach to understanding temporal and thematic shifts in news content.

Specifically, we first average the vector embedding across all news headlines published within each month. For month t, we calculate the novelty of the news relative to the past five months as

$$Novelty_t = 1 - \max_{1 \le j \le 5} \text{Similarity}(\mathbf{e}_t, \mathbf{e}_{t-j}) \tag{1}$$

where $\mathbf{e}_t$ is the average vector representation of the news at month t, and **similarity** is the correlation similarity between two news.

3.5. *Data Description*

We use font-page news on *Wall Street Journal* which is available from Factiva. The news include both headlines and business and finance alerts. The dataset employed in our study stands as one of the most expansive and comprehensive text corpora of business news investigated within the realm of economic literature to date. This dataset encompasses all articles published in front page of *Wall Street Journal* from January 1996 through December 2022, procured from the Dow Jones Historical News Archive. Notably, this dataset represents an unparalleled historical continuum of news articles available for purchase in digital format from Dow Jones & Company, offering an extensive and unparalleled archival record of business news within the financial landscape.

The *Wall Street Journal* stands as an iconic and invaluable resource within the domain of financial research (Baker, Bloom, and Davis, 2016; Manela and Moreira, 2017; Bybee, Kelly, Manela, and Xiu, 2023), wielding a profound influence and playing a pivotal role in shaping the discourse and understanding of financial markets. As a venerable publication renowned for its comprehensive coverage of global financial markets, economic trends, corporate developments, and geopolitical events, the *Wall Street Journal* serves as a cornerstone for academics,

analysts, and practitioners alike seeking authoritative and real-time information.

[Insert Table 1 about here]

Table 1 presents the summary statistics of the *Wall Street Journal* news dataset from January 1996 to December 2022. The dataset encompasses the total count of monthly news articles, categorized into "bad", "neutral", and "good" news based on their anticipated impact on the stock market—signifying the market is going down, uncertain, or going up —according to the ChatGPT-3.5 model's assessment. The average monthly news volume is 260.91, with a standard deviation of 116.12, indicating significant time volatility. A negative skewness of -0.22 suggests a mild leftward tilt in the distribution, while the median of 288 reflects the data's central tendency. The total news range extends from a minimum of 47 to a maximum of 577 items per month, encompassing 84,535 articles over the observed period.

Specifically, the "bad" news segment reports an average of 32.76 articles per month, with a lower variability (standard deviation of 23.93) than the total news count. This category exhibits a positive skewness of 1.15 and a median of 32, indicating a moderate rightward skew in its distribution. The monthly range for "bad" news varies from 0 to 142, totaling 10,613 articles. The "neutral" news, forming the bulk of the dataset, averages 181.75 monthly articles with a standard deviation 74.34. This category shows an almost symmetrical distribution (skewness of -0.07) and a median of 190. The range for "neutral" news stretches from 43 to 422 monthly articles, amounting to 58,888. Lastly, the "good" news category, denoting positive market sentiments, maintains an average of 46.40 monthly articles with a standard deviation 28.25. It exhibits a slight positive skewness (0.10) and a median of 47. The monthly range for "good" news spans from 0 to 121, aggregating 15,034 articles.

[Insert Figure 1 about here]

We define the monthly news ratio as the proportion of good news to the total monthly news count, denoted as NR^G, and the proportion of bad news to the total monthly news count, represented by NR^B. Figure 1 illustrates the temporal progression of these ratios, NR^G and NR^B, through a time series plot.

4. Empirical Results

4.1. Baseline Regression Model

This section explores the impact of the textual information captured by ChatGPT-3.5 on the aggregate stock market. We use a univariate regression model as follows:

$$R_{t+h} = \alpha + \beta\, NR_t^K + \varepsilon_{t+h}\ , \qquad K = B \text{ or } G\ , \tag{2}$$

where R_t denotes the current market excess return on the S&P 500 index at time t for "$h = 0$". This setting aligns with a contemporaneous regression framework. For scenarios where $h > 0$, R_{t+h} represents the average excess returns of the market portfolio from $t + 1$ to $t + h$ (with h being 1, 3, 6, 9, and 12 months), transitioning the equation into a predictive regression. Utilizing ChatGPT-3.5's zero-shot prompt capability, we identify instances of good or bad news from the front-page headlines of the *Wall Street Journal*. The bad news ratio, NR^B, is quantified as the monthly proportion of bad news, while NR^G represents the proportion of good news. The in-sample predictability of both NR^B and NR^G is examined by estimating the regression for the period from January 1996 to December 2022. A critical aspect of our analysis involves assessing the coefficient β ($\hat{\beta}$) in the regression. The null hypothesis presumes that the ChatGPT-estimated textual information lacks predictive power, implying $\beta = 0$ and reducing the regression to a model of constant expected return ($R_{t+h} = \alpha + \varepsilon_{t+h}$). The alternative hypothesis, however, posits that β is non-zero, indicating that GPT-extracted textual information holds significant predictive power for R_{t+h}. For the computation of the corresponding t-statistic for $\hat{\beta}$, we employ the Hodrick (1992) standard error.[2]

[Insert Table 2 about here]

Table 2 presents the estimation results of regression (2). In Panel A, we observe the slope coefficient for the contemporaneous regression on NR^B (when $h = 0$) at −1.03%, with

[2]In analyzing predictability over extended horizons, the Newey and West (1987) may lead to over-rejection in finite samples, particularly when persistent regressors interact with serially correlated errors. The Hodrick's standard error addresses this issue by adopting the moving-average structure of the aggregated error, thereby offering enhanced performance (Ang and Bekaert, 2007).

a Hodrick (1992) t-statistic of -2.70. However, coefficients across various predictive horizons are not statistically significant. This pattern suggests that bad news, as identified by GPT, may not present unexpected insights to human investors, resulting in their immediate reflection in stock returns and thus, diminishing their predictive value. In contrast, Panel B reveals a more gradual market response. Here, the regression coefficient for the contemporaneous relationship between stock market returns and NR^G is 1.02% and is statistically significant. This finding implies that ChatGPT-identified good news positively correlates with the stock market. Notably, this significance persists with coefficients ranging from 0.51% to 0.56% over predictive horizons of 1 to 6 months, though it declines after the 9-month mark. Moreover, the predictive regressions exhibit substantial R^2 values, between 1.37% and 8.52%, which increase with longer horizons, highlighting the significant predictive power of NR^G for stock market returns. These results from Panel B indicate that ChatGPT is capable of extracting aspects of good news not typically discerned by investors, leading to a delayed incorporation of this information into stock prices. The robustness of our findings is further supported by the Newey and West (1987) t-statistic, as presented in Table IA 1 of the Internet Appendix.

Our findings in Table 2 exhibit notable deviations from the observations of Tetlock (2007). In his study, it was noted that daily news pessimism could predict stock market trends, with a reversal observed within a five-day window, whereas news optimism seemed to lack predictive power. This divergence in results probably arises from the differing methodologies employed in textual analysis. Specifically, Tetlock (2007) utilized the Harvard psychology dictionary and word count frequency for text extraction. This traditional approach assigns a static representation to each word, irrespective of its contextual usage. By contrast, our methodology, leveraging Large Language Models (LLMs) and their foundational transformer architecture, affords a more granular and context-sensitive extraction of meanings, operating at both the word and sentence levels (Vaswani et al., 2017; Peters et al., 2018). In subsequent sections, we further juxtapose our approach with the word list methodology proposed by Loughran and McDonald (2011) for identifying positive and negative words, facilitating a comprehensive comparative analysis.

Additionally, the study by Frank and Sanati (2018), utilizing firm-level data, suggests that stock prices have a tendency to overreact to positive news and underreact to negative

news. In contrast, our results demonstrate a more gradual assimilation of positive news into the market, accompanied by an immediate response to negative news. This pattern presents a notable departure from existing theories that advocate for delayed information diffusion (Hong and Stein, 1999; Green, Huang, Wen, and Zhou, 2019; Agrawal, Hacamo, and Hu, 2021), or theories emphasizing limited investor attention (see, for instance, DellaVigna and Pollet, 2009; Hirshleifer, Lim, and Teoh, 2009; Ben-Rephael, Da, and Israelsen, 2017). These theories generally propose a lack of initial reactions in stock returns to new information. An in-depth exploration of this apparent anomaly is reserved for Section **5**, where we delve into potential economic explanations for these observations.

In summary, our analysis reveals that a high proportion of good news, as identified by ChatGPT-3.5, exhibits a significant correlation with both the current and the subsequent returns of the market portfolio extending up to six months. Conversely, the ratio of bad news demonstrates a positive correlation exclusively with current returns, lacking predictive power for future returns. These findings underscore a distinct market response pattern: while the assimilation of positive news identified by ChatGPT-3.5 into market prices is gradual, the response to negative news is immediate.

4.2. Comparisons with Other Methods of Textual Analysis

Our study underscores the robust predictive capability of ChatGPT-3.5 in discerning textual information for future market returns. This efficacy prompts an exploration of whether alternative methods of textual analysis offer comparable predictive insights. A notable approach in this regard is the word lists or bag-of-words methodology, prominently advocated by Loughran and McDonald (2011). They introduced a specialized dictionary of *positive* and *negative* words, specifically designed for financial texts, an approach that has garnered widespread adoption, as evidenced in studies like García, Hu, and Rohrer (2023), Jiang, Lee, Martin, and Zhou (2019), and Cohen, Malloy, and Nguyen (2020).

Employing this dictionary, we categorized front-page news from the *Wall Street Journal* into "good" and "bad" news segments and subsequently computed their respective news ratios. We then re-estimated the regression (2), using the news ratio derived from this word list method as the regressor. The findings, as presented in Table 3, reveal a minimal impact

of good news on stock market returns. In contrast, the ratio of bad news exhibits a significant correlation with current market returns but lacks the ability to predict future returns.

[Insert Table 3 about here]

Our results align with the observations of Tetlock (2007), particularly in underscoring the greater influence of news pessimism compared to optimism on stock returns. However, our findings diverge concerning the forecasting ability of bad news. Tetlock (2007) posited a significant predictive power stemming from media pessimism, a claim not entirely supported by our analysis. A possible reason for this discrepancy could be attributed to the diminished novelty of information extracted via word lists. Since the introduction of this method in 2011, its widespread adoption might have lessened its novelty and, consequently, its predictive value to human investors.

In our subsequent analysis, we evaluate the potential of other LLMs for their abilities to discern textual information and to predict stock market returns. Specifically, we utilized BERT and RoBERTa, developed by *google*, to analyze "good" and "bad"news from the front-page news of the *Wall Street Journal*. Unlike the GPT framework, these models do not readily accommodate natural language instructions (prompts) to differentiate between positive and negative news. To overcome this limitation, we adopted an alternative approach. We began by randomly selecting 300 news articles, then manually classified each into categories: *GOING UP*, *GOING DOWN*, or *UNKNOWN*. Following this classification, we embarked on a process of fine-tuning the parameters of the BERT model. This fine-tuning involved retraining and updating the model's parameters, with a focus on selecting hyper-parameters that optimized accuracy on our validation set.

Table 3 delineates the forecasting outcomes for both the bad and good news ratios as identified by the BERT model. It is observed that the textual information extracted by BERT exerts minimal influence on stock market returns, with a notable exception being the contemporaneous regression that is contingent upon the good news ratio. In a similar vein, the RoBERTa model, as reported in Table IA 2 of the Internet Appendix, demonstrates limited predictive capabilities. These findings align well with the recent scholarly discourse (Brown et al., 2020; Wei et al., 2022; Wei et al., 2022; Zhao et al., 2023), which posits that

larger language models, boasting hundreds of billions or more parameters, exhibit unique "emergent abilities" that are not typically found in smaller models such as BERT and RoBERTa.

[Insert Figure 2 about here]

To elucidate the distribution of words within categories of the good news, bad news, and the unknown news as identified by ChatGPT-3.5, BERT, and the word lists method used in sentiment analysis (Loughran and McDonald, 2011), we conducted a detailed word frequency analysis. This process involved several key steps:

- First, we cleansed the news headlines by removing digits, punctuation, and stop-words using the NLTK package, a widely recognized tool in text analysis. This was followed by lemmatizing the text to retain the base form of each word.[3]
- Next, we computed the word frequency for each headline, subsequently aggregating these frequencies within each category: good news, bad news, and the unknown.
- Finally, we excluded the least frequent words and calculated the relative frequency of remaining words within each category.

Figure 2 illustrates the word clouds for good and bad news as classified by ChatGPT-3.5, the word lists approach of Loughran and McDonald (2011), and the BERT model. Notably, ChatGPT-3.5 adeptly captures context-sensitive words reflective of the financial market. For instance, frequent terms in positive news include "bounce", "notch", "boosting", and "buoy" typically associated with favorable financial or economic conditions. Conversely, the word lists method yields words like "leadership", "beautiful", "improve", and "prospers" for positive news, which, despite being generally affirmative, are less commonly linked to financial contexts. Intriguingly, the BERT model appears less effective in this regard, incorrectly associating words such as "plummet" and "lowest" with positive news, thus contradicting economic intuition.

Overall, this section delves into a comparative analysis of text information predictability as extracted by various methodologies: ChatGPT-3.5, the conventional word lists approach, and other language models with relatively smaller parameter sizes, such as BERT

[3]For more details about NLTK, refer to https://www.nltk.org/.

and RoBERTa. Our investigation reveals that ChatGPT-3.5 demonstrates a notable "emergent" ability. It proficiently extracts significant stock market information that is not as effectively captured by either the traditional word lists approach or the smaller language models. This distinction highlights the advanced capability of ChatGPT-3.5 in processing and interpreting complex textual data relevant to financial markets.

4.3. Comparisons with Macroeconomic Predictors

The content of *Wall Street Journal* may already encompass key aspects of economic fundamentals, raising the question of whether the predictability of NR^G using ChatGPT-3.5 is merely reflective of these underlying economic variables. To explore this possibility, we extend our analysis by incorporating common economic variables as controls in our predictive model. The modified regression model is structured as follows:

$$R_{t+h} = \alpha + \beta\, NR_t^G + \psi\, \mathbf{X}_t + \varepsilon_{t+h}\ , \tag{3}$$

where R_t represents the current market excess return at time t when $h = 0$. For instances where $h > 0$, R_{t+h} denotes the average excess returns of the market portfolio from $t + 1$ to $t + h$, with h varying from 1 to 12 months. Thus, equation (3) functions as a predictive regression. Here, NR^B signifies the bad news ratio as identified by ChatGPT-3.5, while $\mathbf{X}_t$ is a vector of 14 economic variables proposed by Welch and Goyal (2008).[4] A comprehensive description of these variables is available in **Appendix A**.

Using all the macroeconomic variables together in a single regression may result in the potential collinearity issue. Instead, we opted to control for the first four principal components of these variables. The regression outcomes, detailed in Table IA 3 of the Internet Appendix, reveal that the coefficients of NR^G consistently maintain positive values and bear economic significance across various prediction horizons. These results are in alignment with the estimates reported in Table 2, underscoring the robustness of our findings. Crucially, the statistical significance of NR^G remains after the inclusion of common economic variables as controls. This suggests that the information content of NR^G is not merely a reflection of

[4]Data sourced from https://sites.google.com/view/agoyal145

these economic variables but rather provides distinct and valuable insights.

Additionally, we controlled for lagged market returns in our analysis. Given the significant correlation of NR^G with contemporaneous returns, it is imperative to ascertain whether the observed predictability of NR^G is inherently tied to the continuation of stock price. To alleviate this concern, we included the current market return as control variable in the predictive regressions based on NR^G. The detailed results, presented in Table IA 4 of the Internet Appendix, show that the coefficient of NR^G retains its significance, resonating with the findings in Table 2. Conversely, the coefficient for the current return demonstrates insignificance, suggesting that the predictability attributed to NR^G is not merely a manifestation of price momentum. In short, our analysis in this section underscores that the predictive power of NR^G, as extracted by ChatGPT-3.5, is distinct and not merely an overlap with existing fundamental economic variables or stock market trend.

4.4. Robustness Check

In this subsection, we show our results robust to alternative prompts, fine-tuning, and ChatGPT-4. We first use three alternative prompts:

1. "*Forget all previous instructions. You are now a financial expert giving investment advice. I'll give you a news headline, and you need to answer whether this headline is PESSIMISTIC or OPTIMISTIC for the U.S. stock market. Please choose only one option from PESSIMISTIC, OPTIMISTIC, UNKNOWN, and do not provide any additional responses.*"

2. "*Forget all previous instructions. You are now a financial expert giving investment advice. I'll give you a news headline, and you need to answer whether this headline is NEGATIVE or POSITIVE for the U.S. stock market. Please choose only one option from NEGATIVE, POSITIVE, UNKNOWN, and do not provide any additional responses.*"

3. "*Forget all previous instructions. You are now a financial expert giving investment advice. I'll give you a news headline, and you need to answer whether this headline is GOOD or BAD for the U.S. stock market. Please choose only one option from GOOD, BAD, UNKNOWN, and do not provide any additional responses.*"

[Insert Tables 4 and 5 about here]

Utilizing these three distinct prompts, we engaged ChatGPT-3.5 to analyze news articles and compute corresponding news ratios. The regression results based on these newly defined news ratios are tabulated in Table 4, elucidating the outcomes of regression (2). We note that *pessimistic* (or *negative*) news appears to exert minimal impact on stock market returns, as evidenced in Panels A and C. In contrast, *optimistic* (or *positive*) news demonstrates a significant influence on market performance in Panels B and D. Specifically, in the contemporaneous regression, the slope coefficient for the *optimistic* (or *positive*) news ratio registers at 1.09% (0.74%) with a Hodrick (1992) t-statistic of 5.55 (3.31). This significance remains in the predictive regressions, with coefficients ranging from 0.43% to 0.54% in Panel B (and from 0.43% to 0.51% in Panel D). Table IA 5 of the Internet Appendix reports analogous results for the third prompt. These findings collectively attest to the robustness of ChatGPT-3.5's forecasting abilities across a variety of prompts.

Moreover, we employed both ChatGPT-3.5 fine-tuning and ChatGPT-4 to categorize front-page news from the *Wall Street Journal* into good and bad news categories. The forecasting results obtained from these models are delineated in Table 5. Generally, these outcomes are consistent with those detailed in Table 2. Specifically, NR^B, as determined through ChatGPT-3.5 fine-tuning, shows a significant correlation with current returns, yet lacks predictive power for future returns. In contrast, NR^G demonstrates a substantial influence on both current market returns and those over extended periods, with forecasting coefficients varying from 0.34% to 0.80%. This range highlights the pronounced predictive ability of NR^G for the stock market.

The findings from ChatGPT-4 mirror this trend, showing immediate market response to both good and bad news, and a similar pattern in the gradual incorporation of positive news into stock prices, as reflected in Table 2. Notably, our analysis does not indicate a significant performance edge of ChatGPT-4 over ChatGPT-3.5 in stock return prediction. This could be attributed to ChatGPT-4's enhanced proficiency in handling multimodal data, which combines textual and visual elements. Overall, these results affirm the robustness of ChatGPT-3.5's forecasting ability, particularly with respect to fine-tuning and adaptation to

increased model complexity aimed at accommodating multimodal data.

4.5. *Out-of-sample Performance*

This section is dedicated to assessing the out-of-sample return predictability of NR^B and NR^G, as extracted through ChatGPT-3.5. While in-sample analysis facilitates more efficient estimation of parameters and thereby yields more precise return forecasts by leveraging the entirety of available data, studies such as Welch and Goyal (2008) argue for the greater relevance of out-of-sample tests. These tests are deemed crucial in evaluating the actual predictability of returns in a real-time setting, providing a more authentic assessment of the model's predictive power in practical finance scenarios.

We initiate our out-of-sample forecast evaluation with a starting period from January 1996 to December 2005. This period serves as the basis for estimating the monthly predictive regression (2) using NR^B or NR^G, thereby facilitating the generation of our first out-of-sample forecast in January 2006. The forecasted return is articulated as follows:

$$\widehat{R}_{t+1} = \widehat{\alpha}_t + \widehat{\beta}_t \, L_t^I \,, \tag{4}$$

where $\widehat{\alpha}_t$ and $\widehat{\beta}_t$ represent the ordinary least squares (OLS) estimates derived from regression (2). We subsequently engage in a recursive process, continually re-estimating regression (2) and constructing monthly out-of-sample forecasts in accordance with Equation (4). This approach is consistently applied to subsequent periods, extending up to the end of our sample period in December 2022. The selection of the initial in-sample estimation period was strategically made to ensure that the observations were ample for accurately estimating initial parameters, while also allowing for a sufficiently extended out-of-sample period for effective forecast evaluation.[5]

To assess the out-of-sample performance, we implement the widely used Campbell and Thompson (2008)'s R^2_{OS} and Clark and West (2007)'s *MSFE-adjusted* statistical methods. The R^2_{OS} is a measure of the proportional reduction in mean squared forecast error (MSFE) for

[5]Rossi and Timmermann (2010) and Hansen and Timmermann (2012) indicate that out-of-sample tests of predictive ability tend to exhibit improved size properties when the forecast evaluation period constitutes a relatively large proportion of the available sample, as is the case in our analysis.

the predictive regression forecast relative to a benchmark forecast. A positive R^2_{OS} value indicates that the model forecast surpasses the benchmark in terms of MSFE. The benchmark in this context is the average excess return from the start of the sample period up to month t, aligning with the constant expected excess return model delineated in Equation (2) with $\beta = 0$. This implies that returns are not predictable, akin to the canonical random walk model with drift applied to stock prices. To determine whether the predictive regression forecast yields a statistically significant improvement in MSFE, we employ the *MSFE-adjusted* statistic as proposed by Clark and West (2007). This is used to test the null hypothesis $H_0 : R^2_{OS} \leq 0$ against the alternative hypothesis $H_A : R^2_{OS} > 0$, which posits that the historical average MSFE exceeds that of the predictive regression forecast.

[Insert Table 6 about here]

Panel A of Table 6 displays the out-of-sample forecasting results. We observe that while NR^B exhibits a negative R^2_{OS}, the R^2_{OS} for NR^G stands at 1.17% and is statistically significant according to *MSFE-adjusted* statistics. A positive R^2_{OS} suggests that the MSFE for out-of-sample return forecasts based on NR^G is significantly lower than the historical average, indicating substantial economic significance. Given the typically small R^2 in stock return predictions due to the high noise-to-signal ratio, the magnitude of R^2_{OS} for NR^G is notably large. Campbell and Thompson (2008) contend that a monthly R^2_{OS} of 0.5% can have significant economic implications, and our finding of an R^2_{OS} over twice this threshold underscores its substantial economic relevance (Kandel and Stambaugh, 1996). In the following section, we will delve into the economic gains derived from this predictability.

For comparative analysis, we also examine the results for economic variables. According to Rapach, Strauss, and Zhou (2010), the average combination of forecasts from individual economic variables surpasses the performance of a kitchen sink model, using all variables together in a single model, as suggested by Welch and Goyal (2008). Table 6 indicates that this mean combination results in a negative R^2_{OS} of -0.41% in our out-of-sample. Additionally, when comparing with the first four principal components of economic variables discussed in Section **4.3.**, we find an unreported R^2_{OS} of -8.22%, further illustrating the distinct predictive power of NR^G.

[Insert Figure 3 about here]

In light of the pronounced out-of-sample predictability of NR^G, an intriguing line of inquiry pertains to whether this predictability is consistent across the entire sample or confined to specific periods. To investigate this, we adopt the approach suggested by Welch and Goyal (2008), focusing on the temporal evolution of the predictive ability. This involves plotting a time series that represents the difference between the cumulative squared forecast error (CSFE) generated by the historical average benchmark forecast and the CSFE derived from forecasts based on NR^G. A trend characterized by a positive slope in this time-series differential would indicate that NR^G-based forecasts consistently outperform the historical average across various time periods. This graphical representation thereby provides a vivid illustration of the dynamic performance of NR^G as a predictor, over time, further enriching our understanding of its reliability and robustness.

Figure 3 illustrates the difference in CSFE between forecasts based on NR^G and those derived from the mean combination of economic variables. Notably, the curve representing NR^G exhibits a marked upsurge during the periods 2008–2010 and 2021–2022, interspersed with fluctuations across other periods, barring the final year. The overall positive trajectory of the curve signifies a stable predictability of NR^G over time. The recent downward trend may reflect the increasing availability and influence of LLMs, akin to the publication effect posited by McLean and Pontiff (2016). In contrast, the curve associated with the mean combination of economic variables demonstrates a generally negative slope, punctuated only by a transient increase during the 2008 financial crisis. Collectively, Figure 3 underscores the enduring predictive capacity of NR^G across different time frames, thereby affirming its utility in complementing the predictive power inherent in macroeconomic variables.

4.6. Economic Value

In this subsection, we delve into the question of whether the out-of-sample forecasting abilities of NR^B and NR^G, as generated by ChatGPT-3.5, can confer tangible economic benefits to investors. This analysis is particularly pertinent for those contemplating the integration of such forecasting information into their investment strategies, as opposed to

disregarding it. We approach this inquiry from the perspective of asset allocation, aiming to quantify the potential economic gains that might accrue from leveraging the predictive insights offered by NR^B and NR^G.

In alignment with the methodologies advocated by Kandel and Stambaugh (1996), Campbell and Thompson (2008), and Ferreira and Santa-Clara (2011), we consider a mean-variance investor who utilizes return forecasts to make his asset allocation decisions between risky stocks and risk-free bills. Portfolio rebalancing is conducted at the end of each month, with the equity weights in the portfolio determined as per the following equation:

$$w_t = \frac{1}{\gamma} \frac{\widehat{R}_{t+1}}{\widehat{\sigma}^2_{t+1}} , \tag{5}$$

where γ signifies the investor's degree of risk aversion, $\widehat{R}_{t+1}$ represents the out-of-sample forecast of excess stock returns, and $\widehat{\sigma}^2_{t+1}$ is the variance forecast. Consistent with Campbell and Thompson (2008), we presume that investors estimate future stock return variances using a 5-year moving window of past returns. Furthermore, the weight w_t is bounded between 0 and 1.5 to preclude short selling and limit the maximum leverage to 50%.

The investment strategy entails allocating $1 - w_t$ of the portfolio to risk-free bills. Consequently, the realized portfolio return at time $t + 1$, denoted as R^p_{t+1}, is expressed by the following equation:

$$R^p_{t+1} = w_t \, R_{t+1} + R^f_{t+1} , \tag{6}$$

where R_{t+1} represents the excess return of the market portfolio, and R^f_{t+1} is the risk-free return. The certainty equivalent return (CER) of the portfolio is computed as:

$$CER_p = \widehat{\mu}_p - 0.5 \, \gamma \, \widehat{\sigma}^2_p , \tag{7}$$

where $\widehat{\mu}_p$ and $\widehat{\sigma}^2_p$ are the sample mean and variance of the investor's portfolio over the forecast evaluation period, respectively. The CER can be interpreted as the risk-free return that an investor would be willing to accept in lieu of holding a risky portfolio. The CER gain is thus the difference between the CER of an investor utilizing the predictive regression forecast of monthly returns as given by Equation (4) and that of an investor relying on a historical

average forecast. This difference, when multiplied by 12, represents the annual portfolio management fee that an investor might be prepared to pay for access to predictive regression forecasts. Additionally, we calculate the annualized Sharpe ratios of R_t^p to further evaluate investment performance. This approach allows us to directly measure the economic value derived from return predictability.

Panel B of Table 6 delineates the asset allocation results for the out-of-sample period spanning January 2006 to December 2022. For this analysis, we assume a risk aversion coefficient of three. The findings reveal that NR^G achieves a significant CER gain of 4.92%, suggesting that investors might be inclined to pay an annual fee of up to 492 basis points (bps) for access to the predictive regression forecasts based on NR^G. The investment portfolio formulated on the basis of NR^G reports an annualized Sharpe ratio of 0.51, which is substantially higher than the market portfolio's Sharpe ratio of 0.30. This profitability remains considerable even after deducting a proportional transaction cost of 50 basis points, resulting in a net-of-transaction-cost CER gain for NR^G of 3.55%. These asset allocation outcomes demonstrate robustness across different levels of risk aversion, including coefficients of one and five, as detailed in Table IA 6 in the Internet Appendix. In contrast, NR^B yields negative CER gains and lower Sharpe ratios, reflecting its relatively weaker forecasting power compared to NR^G. We also compare with the mean combination of economic variables, and find that it delivers a CER gain of 1.18% with a Sharpe ratio of 0.23. Nevertheless, the economic magnitude of the investment profit is smaller than that of NR^G.

To summarize, our comprehensive analysis underscores the strong forecasting power of the good news ratio, as identified by ChatGPT, for monthly out-of-sample market returns. This predictive ability translates into significant investment profits within the context of asset allocation, thereby indicating considerable economic value for mean-variance investors. Such results suggest the critical importance of the information extracted by GPT models, especially when viewed from the perspective of asset allocation. The implications of these findings are substantial, revealing the potential utility of incorporating ChatGPT-derived insights into investment strategies.

5. Economic Explanations

The findings from our study reveal a distinct pattern in market reactions: a gradual response to good news and a more immediate reaction to bad news, as identified by GPT models. This section explores the potential economic explanations underlying this observed evidence.

5.1. Links to Macroeconomic Conditions

Our initial focus is on elucidating the information content discerned by ChatGPT-3.5. In his seminal research on Intertemporal Capital Asset Pricing Model (ICAPM), Merton (1973) shows that expected excess return of the market portfolio, denoted as $R_{M,t}$, is intrinsically linked to its conditional variance and the conditional covariance with state variable innovations impacting the stochastic investment opportunity set. This relationship is mathematically represented as follows:

$$E(R_{M,t}|\,\Omega_{t-1}) = \beta \cdot Var(R_{M,t}|\,\Omega_{t-1}) + \gamma \cdot Cov(R_{M,t},\, S_t|\,\Omega_{t-1}), \tag{8}$$

where $R_{M,t}$ signifies the excess return of the market portfolio, and S_t represents the innovation in a state variable. The terms $Var(R_{M,t}|\,\Omega_{t-1})$ and $Cov(R_{M,t},\, S_t|\,\Omega_{t-1})$ denote, respectively, the conditional variance of the market excess returns and the conditional covariance between excess market returns and shocks in the investment opportunity set, both conditioned on the information set available up to time $t-1$. Equation (8) posits that expected returns provide compensation to investors for bearing market risk as well as the risk associated with unfavorable shifts in the investment opportunity set.

[Insert Table 7 about here]

In light of recent developments by Bybee, Kelly, and Su (2023), who estimate the state variable from news text using a narrative factor pricing model, we hypothesize that the information extracted by ChatGPT-3.5 may be related to the investment opportunities within the ICAPM framework. To explore this conjecture, we employ several macroeconomic condition proxies, including the Industrial Production Growth (IPG), the CBOE Volatility Index (VIX), the Chicago Fed National Activity Index (CFNAI), the Aruoba-Diebold-Scotti Business Con-

ditions Index (ADSI), the Kansas City Financial Stress Index (KCFSI), Total Non-farm Payroll Growth (Payroll Growth), Smoothed Recession Probability, and Real GDP Growth (GDPG). We regress these proxies on the news ratios as follows:

$$Y_{t+1} = \alpha + \beta\, NR_t^K + \varepsilon_{t+1}\,, \qquad K = B \text{ or } G\,, \tag{9}$$

where NR^K represents the news ratios, either NR^G or NR^B. As evidenced in Table 7, the regression slopes on NR^B are significantly positive for VIX, KCFSI, and SRP, and negative for IPG, CFNAI, and GDPG. This suggests that higher ratio of bad news correlates with heightened market volatility, financial stress, recession probability, and lower industrial production and real GDP growth, indicating powerful ability of ChatGPT to capture the economic downturns. Conversely, NR^G effectively forecasts future macroeconomic conditions, with all regression coefficients being statistically significant. Positive news associates with future high industrial production, real GDP growth, enhanced economic activity, favorable business conditions, employment growth, but lower market volatility and recession probability.

In contrast, when assessing the predictive capabilities of word lists and BERT models (as shown in Table IA 7 of the Internet Appendix), we observe that these methods, although capable of forecasting some macroeconomic variables, exhibit limited economic magnitude and lower t-statistics. This observation provides further insight into their relatively restricted predictive power in stock market contexts, as discussed in Table 3.

In summary, our analysis establishes a significant correlation between our news ratios and future macroeconomic conditions. A higher ratio of good news is indicative of an improving economic state, whereas a higher ratio of bad news suggests deteriorating future economic conditions. This outcome evidences the capability of ChatGPT in capturing pertinent information about the macroeconomic state from news text, through which it influences the aggregate stock market. These findings highlight the potential of advanced language models like ChatGPT in offering insightful macroeconomic forecasts based on their analysis of textual data.

5.2. *Interaction with Economic Activity*

The interplay between news and market response is known to be influenced by the prevailing state of the business cycle. Veronesi (1999) has demonstrated that stock prices tend to underreact to good news during adverse economic times. This phenomenon suggests a nuanced market sensitivity to news contingent upon broader economic conditions. Echoing this perspective, García (2013) have observed that the ability of news content to predict stock returns is more pronounced during recessionary periods. These findings underline the significance of considering the business cycle's phase when assessing the market impact of news, pointing to a potential differential response pattern under varying economic climates.

[Insert Table 8 about here]

To examine the relationship between economic activity and market response to news, we utilize the Chicago Fed National Activity Index (CFNAI) as a proxy for economic conditions, assessing overall economic activity and related inflationary pressures. An indicator variable I_{High} is constructed, assigned a value of one if the current CFNAI exceeds the past five-year sample mean, and zero otherwise. I_{Low} is defined as $1 - I_{High}$. The predictive regression model is formulated as follows:

$$R_{t+h} = \alpha + \beta_1 I_{High} NR_t^G + \beta_2 I_{Low} NR_t^G + \beta_3 I_{High} + \varepsilon_{t+h} , \tag{10}$$

where R_{t+h} represents the average excess return of the market portfolio from $t+1$ to $t+h$, for $h = 1, 3, 6, 9$, and 12 months. The variable NR^G denotes the good news ratio, as identified by GPT-3.5. We primarily focus on the coefficients of $I_{High} \times NR_t^G$ and $I_{Low} \times NR_t^G$. The statistical significance of these coefficients will indicate the predominant periods (high or low economic activity) during which the return predictability of NR_t^G is more pronounced. As depicted in Table 8, our results reveal that while the coefficient for $I_{High} \times NR_t^G$ remains insignificant across various horizons, the coefficient for $I_{Low} \times NR_t^G$ ranges from 0.71% to 0.97%, achieving statistical significance at the 10% level or better. This outcome corroborates the findings of Veronesi (1999) and García (2013), indicating that the return predictability of NR_t^G is markedly more substantial during periods of economic downturns, aligning

with their observations regarding the heightened predictive power of news content during recessions.

5.3. *Interaction with EPU*

In the realm of financial markets, the reaction of investors to public information can be significantly influenced by the degree of information uncertainty. Zhang (2006) posits that in scenarios marked by heightened information uncertainty, investors' tendency to underreact to public information becomes more pronounced. This theory leads to the expectation that the return predictability of good news ratio NR^G, would be more robust during periods characterized by high information uncertainty. The following analysis explores this hypothesis, seeking to understand the predictive strength of NR^G for different levels of economic policy uncertainty (EPU), a proxy of information uncertainty.

[Insert Table 9 about here]

In our analysis, we incorporate the EPU index proposed by Baker, Bloom, and Davis (2016) to assess its interaction with news ratio predictability. An indicator variable, U_{High}, is constructed, which takes the value of one if the current EPU exceeds the past five-year sample mean, and zero otherwise. The counterpart variable, U_{Low}, is defined as $1-U_{High}$. The ensuing regression model is specified to evaluate the impact of these EPU-related indicators on the return predictability of the good news ratio (NR^G):

$$R_{t+h} = \alpha + \beta_1 U_{High} NR_t^G + \beta_2 U_{Low} NR_t^G + \beta_3 U_{High} + \varepsilon_{t+h} . \tag{11}$$

As depicted in Table 9, the estimation results of this regression model (11) highlight the influence of EPU on the return predictability of NR^G. The coefficient for the interaction term $U_{High} \times NR_t^G$ is statistically significant at the 5% level or better, with values ranging from 0.78% to 0.89%. Given that all predictors are standardized with zero mean and unit variance, the economic magnitude of these coefficients implies an increase in returns by 0.78% to 0.89% following a one-standard-deviation increase in NR^G during periods of high EPU. This finding is substantially more pronounced than the results observed during periods of low

EPU, which peak at a maximum of 0.29%. Consistent with the insights of Zhang (2006), our results suggest that high levels of EPU amplify investors' tendency to underreact to good news, thus affecting stock market returns.

In summary, the evidence gathered from our analysis indicates a notable underreaction by investors to good news as identified by the GPT model, particularly during economically challenging times. This underreaction contributes to the observed predictability of the good news ratio (NR^G). Furthermore, this effect of underreaction is found to be more pronounced during periods characterized by high EPU, indicating the influence of broader economic conditions on investor responses to market information. These findings highlight the nuanced ways in which macroeconomic factors and investor behavior interact, as mediated by advanced language models like GPT in the analysis of financial news.

5.4. Interaction with News Novelty

The novelty of news content may also play a crucial role in how investors react to public information. Research by Chan, Jegadeesh, and Lakonishok (1996) indicates that the market's response to newly released information tends to be gradual. Extending this perspective, Daniel, Hirshleifer, and Subrahmanyam (1998) developed a model suggesting that investor overconfidence in personal information can lead to an underreaction to public signals. In a related vein, Tetlock (2011) approached the concept of news novelty by assessing the similarity of current news stories to prior ones regarding the same entity. In line with Tetlock (2011)'s methodology, we adopt a similar approach for evaluating the novelty of economic-relevant news, as delineated in Section ***3.4.***. Here, "economic-relevant news" encompasses stories featuring economic-related keywords, which are detailed in **Appendix B**. This analysis aims to understand how the freshness or staleness of economic news impacts investor behavior and market responses.

[Insert 10 about here]

The regression model constructed to assess the influence of news novelty on market

returns is formulated as follows:

$$R_{t+h} = \alpha + \beta_1 S_{High} NR_t^G + \beta_2 S_{Low} NR_t^G + \beta_3 S_{High} + \varepsilon_{t+h} , \tag{12}$$

where S_{High} represents an indicator variable that takes the value of one when the current news similarity exceeds the past five-year sample mean, and zero otherwise. The counterpart variable S_{Low} is defined as $1 - S_{High}$. As detailed in Table 10, the estimation results for regression (12) reveal notable findings. The coefficient for the interaction term $S_{High} \times NR_t^G$ does not demonstrate statistical significance across different time horizons. Conversely, the coefficient for $S_{Low} \times NR_t^G$ varies between 0.67% and 0.84%, with t-statistics ranging from 2.19 to 3.04. This pattern suggests that the predictability of NR_t^G is more evident when the economic news is particularly novel compared to prior reports. This observation aligns with existing literature which indicates that markets tend to respond gradually to new information. Collectively, our results imply that the unique novelty of news leads to an underreaction from investors to good news, culminating in a more gradual assimilation of this information into market responses.

6. Conclusions

In this paper, we employ ChatGPT to extract good and bad news regarding the stock market from both the news headlines and alerts on the *Wall Street Journal* from 1996 to 2022. Our findings reveal a notable association between a high percentage of identified good news and subsequent market returns at a monthly frequency. This return predictability extends to the next six months and is robust to various prompt setups. Moreover, we find that the text information identified by ChatGPT is more likely to be related to macroeconomic conditions. We also find that the "large" LLMs have the superior ability relative to the small LLMs, like BERT or RoBERTa, and the traditional text analysis method that usually assigns a context-independent representation to each word or sentence.

Furthermore, we find that the LLMs, such as the ChatGPT, have emergent abilities in identifying good news that go beyond the comprehension of human investors. This "over-performance" becomes statistically and economically more significant during economic

downturns, rising economic policy uncertainty, and flourishing news novelty. At the same time, in a sharp contrast, our analysis reveals that human investors do exhibit a relatively efficient capacity to assess, interpret and assimilate bad news. This finding in the efficacy of information digestion between positive and negative news underscores the nuanced nature of investor responses to varying news types that may have wide implications.

In short, our study finds that ChaGPT can predict the market. We uncover the differential abilities of LLMs and human investors in deciphering news information, highlighting the limitations of human comprehension in processing good news especially under distressful market conditions, whereas they are capable of efficiently assessing negative news. These insights contribute to our improved understanding of the interplay between LLMs and human interpretation in the realm of financial market information. Future research is called for to apply LLMs to other financial markets, such as bonds, currencies and commodities, to learn how information processing of ChatGPT and investors differs by asset classes.

Appendix A. Description for Economic Variables

The 14 economic variables of Welch and Goyal (2008) are defined as,

- Dividend-price ratio (log), DP: log of a twelve-month moving sum of dividends paid on the S&P 500 index minus the log of stock prices (S&P 500 index).
- Dividend yield (log), DY: log of a twelve-month moving sum of dividends minus the log of lagged stock prices.
- Earnings-price ratio (log), EP: log of a twelve-month moving sum of earnings on the S&P 500 index minus the log of stock prices.
- Dividend-payout ratio (log), DE: log of a twelve-month moving sum of dividends minus the log of a twelve-month moving sum of earnings.
- Stock return variance, SVAR: sum of squared daily returns on the S&P 500 index.
- Book-to-market ratio, BM: ratio of book value to market value for the Dow Jones Industrial Average.[6]
- Net equity expansion, NTIS: ratio of a twelve-month moving sum of net equity issues by NYSE-listed stocks to the total end-of-year market capitalization of NYSE stocks.
- Treasury bill rate, TBL: interest rate on a three-month Treasury bill (secondary market).
- Long-term yield, LTY: long-term government bond yield.
- Long-term return, LTR: return on long-term government bonds.
- Term spread, TMS: long-term yield minus the Treasury bill rate.
- Default yield spread, DFY: difference between BAA- and AAA-rated corporate bond yields.
- Default return spread, DFR: long-term corporate bond return minus the long-term government bond return.

[6]We compute the logarithm of the book-to-market ratio in the empirical analysis.

- Inflation, INFL: calculated from the CPI for all urban consumers; we use lagged two-month inflation in regression to account for the delay in CPI releases.

Appendix B. Lists of Economic Keywords

The economic-relevant news is the news that includes the following keywords:

Dow Jones, stock exchange, stock prices, stock market, Nasdaq market, Nasdaq stock, security exchange, security price, security market, interest rate, debt market, security, market, economy, fed, bank, finance, monetary

References

Agrawal, A., I. Hacamo, and Z. Hu (2021). Information dispersion across employees and stock returns. *The Review of Financial Studies 34*(10), 4785–4831.

Anastassia, F. (2023). Front-page news: The effect of news positioning on financial markets. Forthcoming in *The Journal of Finance*.

Ang, A. and G. Bekaert (2007). Stock return predictability: Is it there? *The Review of Financial Studies 20*(3), 651–707.

Antweiler, W. and M. Z. Frank (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance 59*(3), 1259–1294.

Baker, S. R., N. Bloom, and S. J. Davis (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics 131*(4), 1593–1636.

Ben-Rephael, A., Z. Da, and R. D. Israelsen (2017). It depends on where you search: Institutional investor attention and underreaction to news. *The Review of Financial Studies 30*(9), 3009–3047.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Bybee, L., B. Kelly, and Y. Su (2023). Narrative asset pricing: Interpretable systematic risk factors from news text. *The Review of Financial Studies 36*(12), 4759–4787.

Bybee, L., B. T. Kelly, A. Manela, and D. Xiu (2023). Business news and business cycles. Forthcoming in *The Journal of Finance*.

Campbell, J. Y. and S. B. Thompson (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies 21*(4), 1509–1531.

Chan, L. K., N. Jegadeesh, and J. Lakonishok (1996). Momentum strategies. *The Journal of Finance 51*(5), 1681–1713.

Chen, Y., B. T. Kelly, and D. Xiu (2023). Expected returns and large language models. Working Paper, Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4416687.

Clark, T. E. and K. D. West (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics 138*(1), 291–311.

Cohen, L., C. Malloy, and Q. Nguyen (2020). Lazy prices. *The Journal of Finance 75*(3), 1371–1415.

Daniel, K., D. Hirshleifer, and A. Subrahmanyam (1998). Investor psychology and security market over and under-reactions. *The Journal of Finance 53*(6), 1839–1885.

DellaVigna, S. and J. M. Pollet (2009). Investor inattention and friday earnings announcements. *The Journal of Finance 64*(2), 709–749.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ferreira, M. A. and P. Santa-Clara (2011). Forecasting stock market returns: The sum of the parts is more than the whole. *Journal of Financial Economics 100*(3), 514–537.

Frank, M. Z. and A. Sanati (2018). How does the stock market absorb shocks? *Journal of Financial Economics 129*(1), 136–153.

García, D. (2013). Sentiment during recessions. *The Journal of Finance 68*(3), 1267–1300.

García, D., X. Hu, and M. Rohrer (2023). The colour of finance words. *Journal of Financial Economics 147*(3), 525–549.

Green, T. C., R. Huang, Q. Wen, and D. Zhou (2019). Crowdsourced employer reviews and stock returns. *Journal of Financial Economics 134*(1), 236–251.

Hansen, P. R. and A. Timmermann (2012). Choice of sample split in out-of-sample forecast evaluation. Working Paper.

Hirshleifer, D., S. S. Lim, and S. H. Teoh (2009). Driven to distraction: Extraneous events and underreaction to earnings news. *The Journal of Finance 64*(5), 2289–2325.

Hodrick, R. J. (1992). Dividend yields and expected stock returns: Alternative procedures for inference and measurement. *The Review of Financial Studies 5*(3), 357–386.

Hong, H. and J. C. Stein (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of Finance 54*(6), 2143–2184.

Jiang, F., J. Lee, X. Martin, and G. Zhou (2019). Manager sentiment and stock returns. *Journal of Financial Economics 132*(1), 126–149.

Kandel, S. and R. F. Stambaugh (1996). On the predictability of stock returns: An asset-allocation perspective. *The Journal of Finance 51*(2), 385–424.

Kaplan, J., S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics 45*(2-3), 221–247.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lopez-Lira, A. and Y. Tang (2023). Can chatgpt forecast stock price movements? Return predictability and large language models. Working Paper, Availabe at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4412788.

Loughran, T. and B. McDonald (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance 66*(1), 35–65.

Loughran, T. and B. McDonald (2020). Textual analysis in finance. *Annual Review of Financial Economics 12*, 357–375.

Manela, A. and A. Moreira (2017). News implied volatility and disaster concerns. *Journal of Financial Economics 123*(1), 137–162.

McLean, R. D. and J. Pontiff (2016). Does academic research destroy stock return predictability? *The Journal of Finance 71*(1), 5–32.

Merton, R. C. (1973). An intertemporal capital asset pricing model. *Econometrica 41*(5), 867–887.

Newey, W. K. and K. D. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*(3), 703–708.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer (2018, June). Deep contextualized word representations. In M. Walker, H. Ji, and A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, pp. 2227–2237. Association for Computational Linguistics.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever (2019). Language models are unsupervised multitask learners.

Rapach, D. E., J. K. Strauss, and G. Zhou (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies 23*(2), 821–862.

Rogers, A., O. Kovaleva, and A. Rumshisky (2021). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics 8*, 842–866.

Rossi, A. G. and A. Timmermann (2010). What is the shape of the risk-return relation? In *AFA 2010 Atlanta Meetings Paper*.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance 62*(3), 1139–1168.

Tetlock, P. C. (2011). All the news that's fit to reprint: Do investors react to stale information? *The Review of Financial Studies 24*(5), 1481–1512.

Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance 63*(3), 1437–1467.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). Attention is all you need. *Advances in neural information processing systems 30*.

Veronesi, P. (1999). Stock market overreactions to bad news in good times: A rational expectations equilibrium model. *The Review of Financial Studies 12*(5), 975–1007.

Wei, J., Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus (2022, June). Emergent Abilities of Large Language Models. *arXiv e-prints*, arXiv:2206.07682.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou (2022). Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Welch, I. and A. Goyal (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies 21*(4), 1455–1508.

Zhang, X. F. (2006). Information uncertainty and stock returns. *The Journal of Finance 61*(1), 105–137.

Zhao, W. X., K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Figure 1: Time Series Plots of News Ratios

This figure depicts the time series of monthly good news ratio (NR^G) and bad news ratio (NR^B) from January 1996 to December 2022. NR^G (NR^B) is defined as the monthly proportion of good (bad) news identified by ChatGPT-3.5. It answers whether the input news means *GOING UP* (*GOING DOWN*) for stock market. The vertical bars correspond to the NBER-dated recessions.

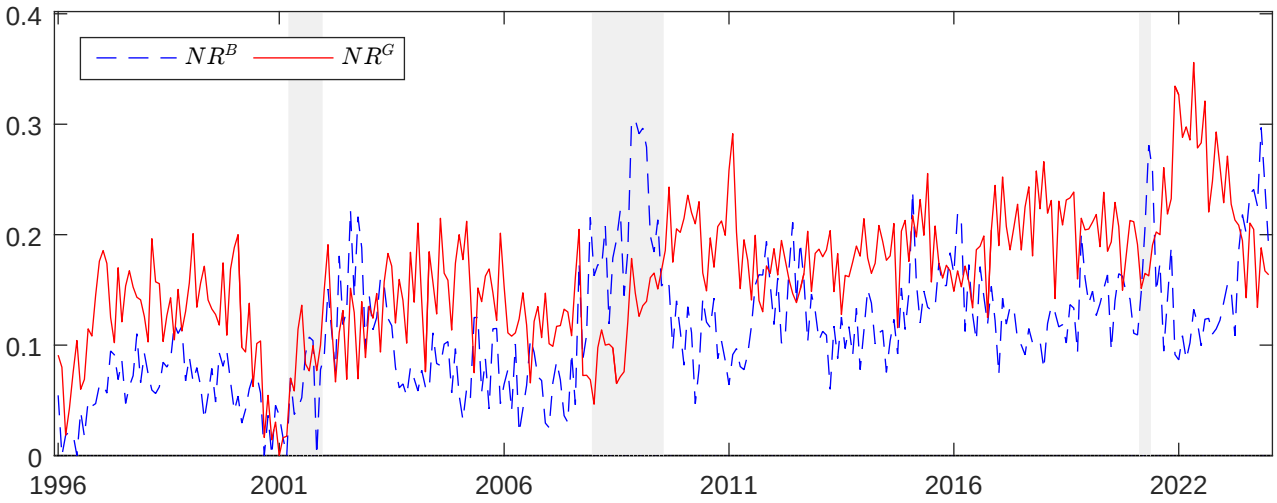

Figure 2: Word Cloud Identified by Various Language Models

This figure shows the word distribution for the good news and bad news identified by various methods. The two sub-figures of the first row shows word distribution for the good news and bad news, respectively, identified by the ChatGPT-3.5. Similarly, the second and third rows show the word cloud for the good news and bad news identified by the word lists proposed by Loughran and McDonald (2011) and the BERT model, respectively.

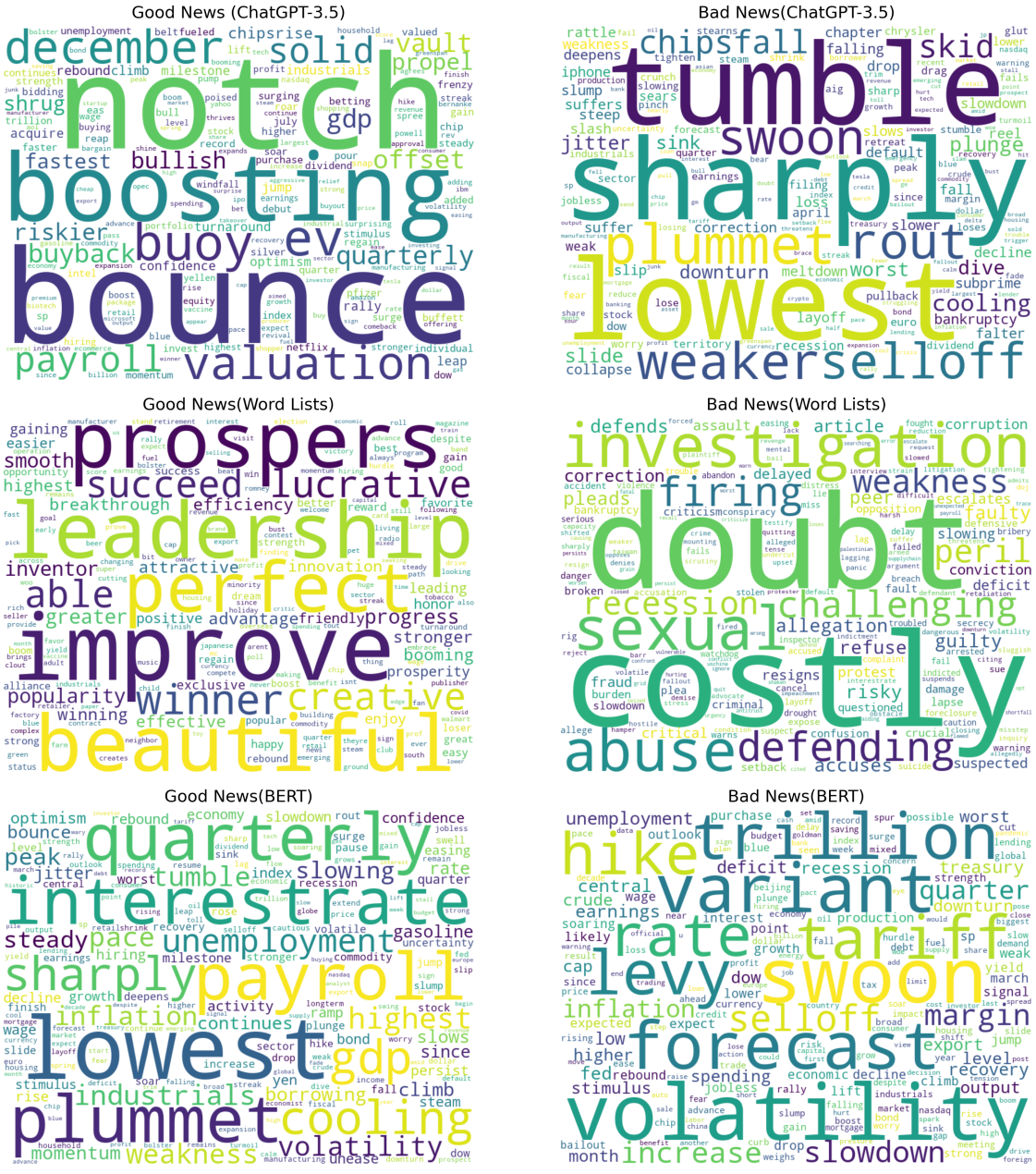

Figure 3: Differences in Cumulative Squared Forecast Errors

This figure plots the difference between the cumulative squared forecast error (CSFE) generated by the historical average benchmark forecast and the CSFE derived from forecasts based on good news ratio, NR^G, which is the monthly proportion of good news identified by ChatGPT-3.5. It answers whether the input news means *GOING UP* for stock market. As a comparison, the figure also depicts the difference in CSFE for the mean combination of forecasts based on the 14 economic variables of Welch and Goyal (2008). The out-of-sample period spans from January 2006 to December 2022. Grey shadow bars denote NBER recessions.

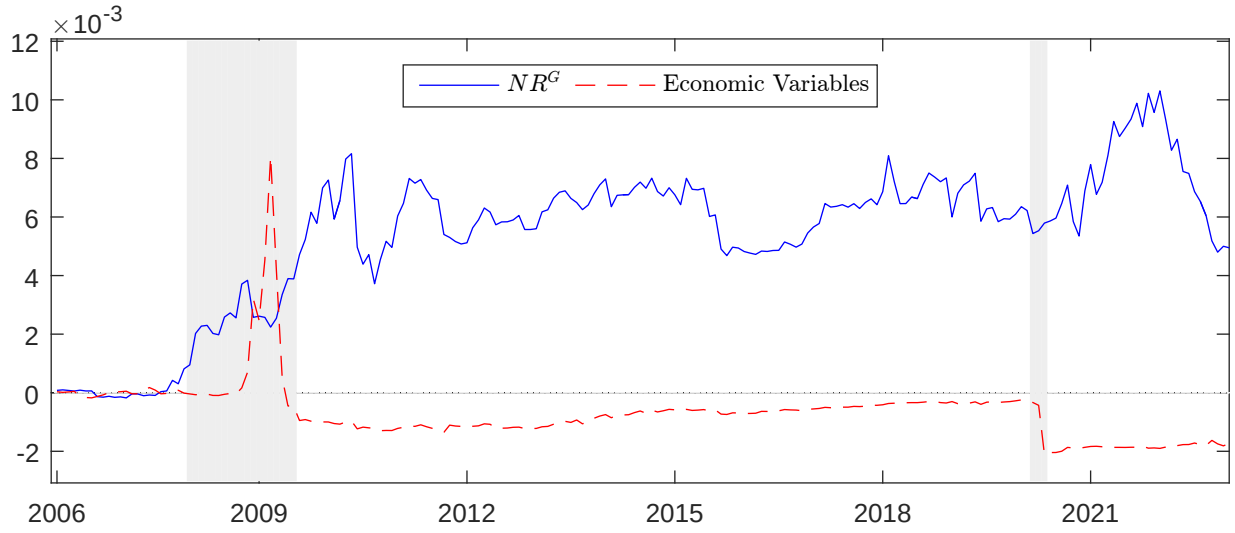

Table 1: Summary Statistics of *Wall Street Journal* News

This table reports the mean, standard deviation (Std. Dev.), skewness (Skew.), median, minimum (min.) and maximum (Max.) of the total monthly news, monthly bad news, monthly neutral news, and monthly good news. The good, neutral, or bad news is identified by ChatGPT-3.5. It answers whether the input news means *GOING UP*, *GOING DOWN*, or *UNKNOWN* for stock market. In the last column, we present the number of news in our sample from January 1996 to December 2022.

	Mean	Std. Dev.	Skew.	Median	Min.	Max.	No. of News
Total News	260.91	116.12	−0.22	288	47	577	84535
Bad News	32.76	23.93	1.15	32	0	142	10613
Neutral News	181.75	74.34	−0.07	190	43	422	58888
Good News	46.40	28.25	0.10	47	0	121	15034

Table 2: Forecasting Results for ChatGPT-3.5

This table reports estimation results of the following regression,

$$R_{t+h} = \alpha + \beta\, NR_t^K + \varepsilon_{t+h}\,, \qquad K = B \text{ or } G\,,$$

where R_t denotes the current market excess return on the S&P 500 index at time t for "$h = 0$". This setting aligns with a contemporaneous regression framework. For scenarios where $h > 0$, R_{t+h} represents the average excess returns of the market portfolio from $t + 1$ to $t + h$ (with h being 1, 3, 6, 9, and 12 months), transitioning the equation into a predictive regression. NR^K represents news ratios. Specifically, NR^B is the monthly proportion of bad news and NR^G represents the proportion of good news. The good or bad news is identified by ChatGPT-3.5. It answers whether the input news means *GOING UP* or *GOING DOWN* for the stock market. Reported are the regression slopes and R^2s in percentage form. Also reported are the Hodrick (1992) t-statistics. All the forecast variables are standardized to have a zero mean and unit variance. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from January 1996 to December 2022.

	β (%)	Hodrick-t	R^2 (%)
Panel A: Results for NR^B			
$h = 0$	-1.03^{***}	-2.70	5.17
$h = 1$	0.05	0.14	0.01
$h = 3$	0.06	0.18	0.05
$h = 6$	0.14	0.47	0.48
$h = 9$	0.21	0.76	1.57
$h = 12$	0.25	0.94	2.65
Panel B: Results for NR^G			
$h = 0$	1.02^{***}	5.30	5.07
$h = 1$	0.53^{**}	2.22	1.37
$h = 3$	0.56^{**}	2.34	4.60
$h = 6$	0.51^{*}	1.84	6.91
$h = 9$	0.44	1.51	7.46
$h = 12$	0.42	1.48	8.52

Table 3: Comparisons with Alternative Textual Analysis Methods

This table reports estimation results of the following regression,

$$R_{t+h} = \alpha + \beta\, NR_t^K + \varepsilon_{t+h}\,, \qquad K = B \text{ or } G\,,$$

where R_t denotes the current market excess return on the S&P 500 index at time t for "$h = 0$". This setting aligns with a contemporaneous regression framework. For scenarios where $h > 0$, R_{t+h} represents the average excess returns of the market portfolio from $t+1$ to $t+h$ (with h being 1, 3, 6, 9, and 12 months), transitioning the equation into a predictive regression. NR^K represents news ratios. Specifically, NR^B is the monthly proportion of bad news and NR^G represents the proportion of good news. We identify the good or bad news by using the method of word lists proposed by Loughran and McDonald (2011) or using BERT. Reported are the regression slopes and R^2s in percentage form. Also reported are the Hodrick (1992) t-statistics. All the forecast variables are standardized to have a zero mean and unit variance. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from January 1996 to December 2022.

	Panel A: Results for NR^B			Panel B: Results for NR^G		
	β (%)	Hodrick-t	R^2 (%)	β (%)	Hodrick-t	R^2 (%)
Word Lists						
$h = 0$	-0.77^{***}	-2.68	2.90	0.24	1.05	0.27
$h = 1$	-0.17	-0.47	0.14	-0.23	-0.87	0.27
$h = 3$	-0.14	-0.56	0.28	-0.14	-0.77	0.27
$h = 6$	0.05	0.27	0.06	-0.09	-0.74	0.20
$h = 9$	0.09	0.46	0.31	-0.14	-1.38	0.70
$h = 12$	0.11	0.49	0.56	-0.11	-1.36	0.60
Bert						
$h = 0$	-0.38	-1.03	0.71	-0.55^{*}	-1.87	1.49
$h = 1$	0.10	0.36	0.05	0.32	1.00	0.49
$h = 3$	0.13	0.71	0.25	0.22	0.82	0.72
$h = 6$	0.18	0.86	0.81	0.23	0.82	1.38
$h = 9$	0.09	0.43	0.29	0.22	0.84	1.68
$h = 12$	0.10	0.53	0.41	0.32	1.32	4.12

Table 4: Results for Alternative Prompts

This table reports estimation results of the following regression,

$$R_{t+h} = \alpha + \beta\, NR_t^K + \varepsilon_{t+h}\,, \qquad K = B \text{ or } G\,,$$

where R_t denotes the current market excess return on the S&P 500 index at time t for "$h = 0$". This setting aligns with a contemporaneous regression framework. For scenarios where $h > 0$, R_{t+h} represents the average excess returns of the market portfolio from $t+1$ to $t+h$ (with h being 1, 3, 6, 9, and 12 months), transitioning the equation into a predictive regression. NR^K represents news ratios. Specifically, NR^B is the monthly proportion of bad news and NR^G represents the proportion of good news. The good or bad news is identified by ChatGPT-3.5. It answers whether the input news is *OPTIMISTIC* (*PESSIMISTIC*) news for the stock market, or is *POSITIVE* (*NEGATIVE*) news for the stock market. Reported are the regression slopes and R^2s in percentage form. Also reported are the Hodrick (1992) t-statistics. All the forecast variables are standardized to have a zero mean and unit variance. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from January 1996 to December 2022.

	β (%)	Hodrick-t	R^2 (%)	β (%)	Hodrick-t	R^2 (%)
	Panel A: *Pessimistic* News			Panel B: *Optimistic* News		
$h = 0$	-0.84^{**}	-2.34	3.46	1.09^{***}	5.55	5.76
$h = 1$	0.03	0.10	0.01	0.43^{*}	1.92	0.88
$h = 3$	0.07	0.20	0.06	0.54^{**}	2.52	4.42
$h = 6$	0.11	0.36	0.29	0.52^{**}	2.07	7.37
$h = 9$	0.18	0.62	1.13	0.49^{*}	1.80	9.30
$h = 12$	0.25	0.90	2.80	0.47^{*}	1.71	10.99
	Panel C: *Negative* News			Panel D: *Positive* News		
$h = 0$	-0.49	-1.57	1.15	0.74^{***}	3.31	2.65
$h = 1$	-0.04	-0.13	0.01	0.43^{*}	1.84	0.88
$h = 3$	0.03	0.10	0.01	0.45^{**}	1.97	3.00
$h = 6$	0.10	0.39	0.27	0.51^{**}	1.96	7.04
$h = 9$	0.20	0.74	1.45	0.47	1.62	8.41
$h = 12$	0.28	1.00	3.72	0.45	1.50	10.00

Table 5: Results for ChatGPT-3.5 Fine–tuning and ChatGPT-4

This table reports estimation results of the following regression,

$$R_{t+h} = \alpha + \beta\, NR_t^K + \varepsilon_{t+h}\,, \qquad K = B \text{ or } G\,,$$

where R_t denotes the current market excess return on the S&P 500 index at time t for "$h = 0$". This setting aligns with a contemporaneous regression framework. For scenarios where $h > 0$, R_{t+h} represents the average excess returns of the market portfolio from $t+1$ to $t+h$ (with h being 1, 3, 6, 9, and 12 months), transitioning the equation into a predictive regression. NR^K represents news ratios. Specifically, NR^B is the monthly proportion of bad news and NR^G represents the proportion of good news. The news is identified by ChatGPT-3.5 fine–tuning and ChatGPT-4, respectively. It answers whether the input news means *GOING UP* or *GOING DOWN* for the stock market. Reported are the regression slopes and R^2s in percentage form. Also reported are the Hodrick (1992) t-statistics. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from January 1996 to December 2022.

	Panel A: Results for NR^B			Panel B: Results for NR^G		
	β (%)	Hodrick-t	R^2 (%)	β (%)	Hodrick-t	R^2 (%)
ChatGPT-3.5 Fine–tuning						
$h = 0$	-1.38^{***}	-4.69	9.20	0.80^{***}	4.04	3.11
$h = 1$	0.14	0.48	0.09	0.50^{**}	2.25	1.23
$h = 3$	0.26	0.82	0.93	0.43^{**}	2.44	2.71
$h = 6$	0.21	0.71	1.05	0.46^{**}	2.57	5.83
$h = 9$	0.17	0.57	0.92	0.39^{**}	2.09	5.74
$h = 12$	0.21	0.81	1.72	0.34^{**}	2.16	5.75
ChatGPT-4						
$h = 0$	-1.08^{***}	-3.68	5.62	0.90^{***}	4.20	3.90
$h = 1$	-0.14	-0.40	0.10	0.37	1.56	0.66
$h = 3$	-0.09	-0.25	0.13	0.36^{*}	1.83	1.90
$h = 6$	-0.07	-0.22	0.14	0.36^{*}	1.83	3.53
$h = 9$	-0.01	-0.03	0.00	0.32	1.61	3.95
$h = 12$	0.09	0.32	0.38	0.28	1.60	3.81

Table 6: Out-of-sample Results for ChatGPT-3.5 and Economic Values

Panel A reports the out-of-sample R^2_{OS}'s, *MSFE-adjusted* statistics, and corresponding p-values for predicting the monthly stock market returns based on news ratios (NR^B or NR^G). NR^G (NR^B) is the monthly proportion of good (bad) news identified by ChatGPT-3.5. It answers whether the input news means *GOING UP* or *GOING DOWN* for stock market. In the last row, we also present the results for mean combination of the forecasts based on 14 economic variables proposed by Welch and Goyal (2008). The regression slopes are estimated recursively using the data available at the forecast formation time t. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. Panel B presents the CER gains in percentage points and annualized Sharpe ratio for a mean-variance investor who allocates assets between the market portfolio and risk-free bills, with a risk-aversion coefficient of three. A proportional transaction cost (TC) of 50bp is also considered. The out-of-sample evaluation period is from January 2006 to December 2022.

Panel A: Out-of-sample Forecasting Results

	R^2_{OS} (%)	*MSFE-adjusted*	p-value
NR^G	1.17**	2.17	0.01
NR^B	−2.55	−1.07	0.86
Economic Variables	−0.41	0.07	0.47

Panel B: Asset Allocation Results

	CER gain (%)	CER gain (%), TC=50bp	Sharpe Ratio
NR^G	4.92	3.55	0.51
NR^B	−3.23	−3.90	−0.07
Economic Variables	1.18	0.87	0.23

Table 7: Forecasting Macroeconomic Conditions

This table presents the results of following regression,

$$Y_{t+1} = \alpha + \beta\, NR_t^K + \varepsilon_{t+1}\,, \qquad K = B \text{ or } G\,,$$

where Y_{t+1} represents macroeconomic condition variables at future time $t+1$, including the industry production growth (IPG), the VIX of CBOE, the CFNAI, the Aruoba-Diebold-Scotti Business Conditions Index (ADSI), the Kansas City Financial Stress Index (KCFSI), the total non-farm payroll growth (Payroll Growth), the smoothed recession probability, and the real GDP growth (GDPG). NR^K represents news ratios. Specifically, NR^B is the monthly proportion of bad news and NR^G represents the proportion of good news. The good or bad news is identify by ChatGPT-3.5. It answers whether the input news means *GOING UP* or *GOING DOWN* for the stock market. Reported are slope estimates (β) and R^2s in percentage form, and also the Newey and West (1987) t-statistics. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from January 1996 to December 2022.

	Panel A: Bad News			Panel B: Good News		
	β	NW-t	R^2 (%)	β	NW-t	R^2 (%)
IPG	-0.18^{**}	-1.97	3.33	0.09^{**}	2.35	0.89
VIX	0.33^{***}	3.80	11.06	-0.14^{**}	-2.55	2.05
CFNAI	-0.17^{*}	-1.69	2.84	0.12^{***}	4.30	1.38
ADSI	-0.13	-1.37	1.56	0.10^{***}	3.66	1.00
KCFSI	0.42^{***}	3.30	17.79	-0.31^{***}	-5.96	9.77
Payroll Growth	-0.09	-1.03	0.85	0.07^{***}	4.26	0.51
SRP	0.39^{***}	3.15	15.13	-0.23^{***}	-3.64	5.46
GDPG	-0.33^{***}	-2.84	10.66	0.10^{*}	1.91	0.97

Table 8: Interaction between Good News Ratio and Economic Activity

This table reports estimation results of the following regression,

$$R_{t+h} = \alpha + \beta_1 I_{High} NR_t^G + \beta_2 I_{Low} NR_t^G + \beta_3 I_{High} + \varepsilon_{t+h} ,$$

where R_{t+h} is the average excess return of market portfolio from $t+1$ to $t+h$, $h = 1$, 3, 6, 9, and 12 months. NR^G is the monthly proportion of good news identified by ChatGPT-3.5. It answers whether the input news means *GOING UP* for the stock market. I_{High} is an indicator variable which equals one if current Chicago Fed National Activity Index (CFNAI) exceeds the past five-year sample mean and zero otherwise. The counterpart variable I_{Low} equals to $1 - I_{High}$. All forecasting variables are standardized to have a zero mean and unit variance. Reported are the regression slopes and R^2s in percentage form. Brackets below the slope estimates report the Hodrick (1992) t-statistics. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from January 1996 to December 2022.

	$h = 1$	$h = 3$	$h = 6$	$h = 9$	$h = 12$
$I_{High} \times NR^G$	0.12	0.17	−0.03	−0.08	−0.06
	[0.44]	[0.68]	[−0.12]	[−0.38]	[−0.38]
$I_{Low} \times NR^G$	0.71*	0.83***	0.97***	0.93***	0.84***
	[1.86]	[2.91]	[3.35]	[3.08]	[3.08]
I_{High}	0.90	0.55	0.53*	0.42	0.44*
	[1.58]	[1.49]	[1.73]	[1.40]	[1.74]
R^2 (%)	1.79	6.36	14.42	17.64	19.64

Table 9: Interaction between Good News Ratio and EPU

This table reports estimation results of the following regression,

$$R_{t+h} = \alpha + \beta_1 U_{High} NR_t^G + \beta_2 U_{Low} NR_t^G + \beta_3 U_{High} + \varepsilon_{t+h} ,$$

where R_{t+h} is the average excess return of market portfolio from $t+1$ to $t+h$, $h = 1, 3, 6, 9$, and 12 months. NR^G is the monthly proportion of good news identified by ChatGPT-3.5. It answers whether the input news means *GOING UP* for the stock market. U_{High} is an indicator variable which equals one if current Economic Policy Uncertainty (EPU) exceeds the past five-year sample mean and zero otherwise. The counterpart variable U_{Low} equals to $1 - U_{High}$. All forecasting variables are standardized to have a zero mean and unit variance. Reported are the regression slopes and R^2s in percentage form. Brackets below the slope estimates report the Hodrick (1992) t-statistics. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from January 1996 to December 2022.

	$h = 1$	$h = 3$	$h = 6$	$h = 9$	$h = 12$
$U_{High} \times NR^G$	0.78**	0.81***	0.84***	0.89***	0.85***
	[2.20]	[3.11]	[2.90]	[3.02]	[2.90]
$U_{Low} \times NR^G$	0.26	0.29	0.15	−0.05	−0.05
	[0.99]	[1.07]	[0.59]	[−0.23]	[−0.25]
U_{High}	0.13	0.26	0.02	−0.07	0.09
	[0.28]	[0.72]	[0.05]	[−0.23]	[0.39]
R^2 (%)	0.79	4.95	9.35	15.15	17.77

Table 10: Interaction between Good News Ratio and News Similarity

This table reports estimation results of the following regression,

$$R_{t+h} = \alpha + \beta_1 S_{High} NR_t^G + \beta_2 S_{Low} NR_t^G + \beta_3 S_{High} + \varepsilon_{t+h} ,$$

where R_{t+h} is the average excess return of market portfolio from $t+1$ to $t+h$, $h = 1$, 3, 6, 9, and 12 months. NR^G is the monthly proportion of good news identified by ChatGPT-3.5. It answers whether the input news means *GOING UP* for the stock market. S_{High} is an indicator variable which equals one if current similarity of economic news exceeds the past five-year sample mean and zero otherwise. The counterpart variable S_{Low} equals to $1 - S_{High}$. The economic news refers to news that contain economic-relevant keywords listed in Appendix B. All forecasting variables are standardized to have a zero mean and unit variance. Reported are the regression slopes and R^2s in percentage form. Brackets below the slope estimates report the Hodrick (1992) t-statistics. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from January 1996 to December 2022.

	$h = 1$	$h = 3$	$h = 6$	$h = 9$	$h = 12$
$S_{High} \times NR^G$	0.24	0.44	0.34	0.26	0.22
	[0.75]	[1.37]	[0.88]	[0.72]	[0.71]
$S_{Low} \times NR^G$	0.84**	0.71**	0.71***	0.67***	0.68***
	[2.35]	[2.19]	[2.97]	[2.83]	[3.04]
S_{High}	0.65	−0.12	0.02	−0.12	−0.20
	[1.19]	[−0.27]	[0.06]	[−0.41]	[−1.35]
R^2(%)	1.38	4.02	6.96	8.23	10.61

Internet Appendix for

ChatGPT, Stock Market Predictability and Links to the Macroeconomy

This Internet Appendix reports the results for supplementary and robustness tests:

Table IA 1: Robust Check for Table 2

Table IA 2: Comparisons with Alternative LLMs

Table IA 3: Comparisons with Economic Variables

Table IA 4: Control for Lagged Return

Table IA 5: Additional Results in Table 4

Table IA 6: Supplementary Results in Table 6

Table IA 7: Additional Results in Table 7

Table IA 1. Forecasting Results for NW-t

This table reports estimation results of the following regression,

$$R_{t+h} = \alpha + \beta\, NR_t^K + \varepsilon_{t+h}\,, \qquad K = B \text{ or } G\,,$$

where R_t denotes the current market excess return on the S&P 500 index at time t for "$h = 0$". This setting aligns with a contemporaneous regression framework. For scenarios where $h > 0$, R_{t+h} represents the average excess returns of the market portfolio from $t+1$ to $t+h$ (with h being 1, 3, 6, 9, and 12 months), transitioning the equation into a predictive regression. NR^K represents news ratios. Specifically, NR^B is the monthly proportion of bad news and NR^G represents the proportion of good news. The good or bad news is identified by ChatGPT-3.5. It answers whether the input news means *GOING UP* or *GOING DOWN* for the stock market. Reported are the regression slopes and R^2s in percentage form. Also reported are the Newey and West (1987) t-statistics (NW-t). All the forecast variables are standardized to have a zero mean and unit variance. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from January 1996 to December 2022.

	β (%)	NW-t	R^2 (%)
Panel A: Results for NR^B			
$h = 0$	-1.03^{***}	-2.77	5.17
$h = 1$	0.05	0.14	0.01
$h = 3$	0.06	0.21	0.05
$h = 6$	0.14	0.54	0.48
$h = 9$	0.21	0.85	1.57
$h = 12$	0.25	1.03	2.65
Panel B: Results for NR^G			
$h = 0$	1.02^{***}	4.78	5.07
$h = 1$	0.53^{**}	2.21	1.37
$h = 3$	0.56^{**}	2.66	4.60
$h = 6$	0.51^{*}	2.21	6.91
$h = 9$	0.44^{*}	1.83	7.46
$h = 12$	0.42^{*}	1.74	8.52

Table IA 2. Forecasting Results for RoBERTa

This table reports estimation results of the following regression,

$$R_{t+h} = \alpha + \beta\, NR_t^K + \varepsilon_{t+h}\,, \qquad K = B \text{ or } G\,,$$

where R_t denotes the current market excess return on the S&P 500 index at time t for "$h = 0$". This setting aligns with a contemporaneous regression framework. For scenarios where $h > 0$, R_{t+h} represents the average excess returns of the market portfolio from $t + 1$ to $t + h$ (with h being 1, 3, 6, 9, and 12 months), transitioning the equation into a predictive regression. NR^K represents news ratios. Specifically, NR^B is the monthly proportion of bad news and NR^G represents the proportion of good news. The good or bad news is identified by RoBERTa. It answers whether the input news means *GOING UP* or *GOING DOWN* for the stock market. Reported are the regression slopes and R^2s in percentage form. Also reported are the Hodrick (1992) t-statistics. All the forecast variables are standardized to have a zero mean and unit variance. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from January 1996 to December 2022.

	Panel A: Results for NR^B			Panel B: Results for NR^G		
	β (%)	Hodrick-t	R^2 (%)	β (%)	Hodrick-t	R^2 (%)
$h = 0$	-0.61^*	-1.94	1.84	-0.16	-0.63	0.12
$h = 1$	0.36	1.32	0.64	0.17	0.63	0.13
$h = 3$	0.24	0.95	0.86	0.32^*	1.81	1.44
$h = 6$	0.22	0.89	1.24	0.25	1.60	1.62
$h = 9$	0.16	0.62	0.89	0.16	1.02	0.90
$h = 12$	0.20	0.78	1.63	0.17	1.26	1.29

Table IA 3. Comparisons with Economic Variables

This table reports estimation results of the following regression,

$$R_{t+h} = \alpha + \beta_1 NR_t^G + \beta_2 PC1_t + \beta_3 PC2_t + \beta_4 PC3_t + \beta_5 PC4_t + \varepsilon_{t+h} ,$$

where R_t denotes the current market excess return on the S&P 500 index at time t for "$h = 0$". This setting aligns with a contemporaneous regression framework. For scenarios where $h > 0$, R_{t+h} represents the average excess returns of the market portfolio from $t+1$ to $t+h$ (with h being 1, 3, 6, 9, and 12 months), transitioning the equation into a predictive regression. NR^G represents the proportion of good news identified by ChatGPT-3.5. It answers whether the input news means *GOING UP* for the stock market. PC1–PC4 are the first four principal components of the 14 economic variables proposed by Welch and Goyal (2008). Reported are the regression slopes and R^2s in percentage form. Brackets below the slope estimates report the Hodrick (1992) t-statistics. All the forecast variables are standardized to have a zero mean and unit variance. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from January 1996 to December 2022.

	β_1	β_2	β_3	β_4	β_5	R^2 (%)
$h = 0$	0.64^{***}	0.29	−0.06	-1.84^{***}	0.23	21.51
	[3.31]	[0.93]	[−0.22]	[−7.21]	[1.07]	
$h = 1$	0.44^{*}	−0.20	0.26	−0.01	−0.23	2.08
	[1.81]	[−0.48]	[0.75]	[−0.02]	[−0.86]	
$h = 3$	0.51^{**}	−0.23	0.11	0.03	−0.25	6.41
	[2.23]	[−0.60]	[0.38]	[0.09]	[−1.09]	
$h = 6$	0.48^{*}	−0.33	0.04	0.12	−0.30	12.63
	[1.67]	[−1.17]	[0.17]	[0.74]	[−1.26]	
$h = 9$	0.39	-0.40^{*}	0.07	0.11	−0.31	17.38
	[1.28]	[−1.78]	[0.33]	[0.97]	[−1.27]	
$h = 12$	0.33	-0.43^{**}	0.10	0.07	−0.31	22.39
	[1.25]	[−2.28]	[0.54]	[0.73]	[−1.32]	

Table IA 4. Control for Lagged Return

This table reports estimation results of the following regression,

$$R_{t+h} = \alpha + \beta\, NR_t^G + \psi\, R_t + \varepsilon_{t+h}\ ,$$

where R_{t+h} is the average excess returns of market portfolio from $t+1$ to $t+h$, $h = 1, 3, 6, 9$, and 12 months. NR^G is the monthly proportion of good news identified by ChatGPT-3.5. It answers whether the input news means *GOING UP* for the stock market. R_t is the current market excess return at time t, which is added as a control variable. Reported are the regression slopes and R^2s in percentage form. Also reported are the Hodrick (1992) t-statistics. All the forecast variables are standardized to have a zero mean and unit variance. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from January 1996 to December 2022.

	β (%)	Hodrick-t	ψ (%)	Hodrick-t	R^2 (%)
$h = 1$	0.54**	2.27	−0.03	−0.12	1.37
$h = 3$	0.57**	2.50	−0.06	−0.33	4.65
$h = 6$	0.52*	1.90	−0.06	−0.68	7.01
$h = 9$	0.44	1.50	−0.02	−0.23	7.47
$h = 12$	0.42	1.49	−0.02	−0.33	8.54

Table IA 5. Additional Results of Alternative Prompts

This table reports estimation results of the following regression,

$$R_{t+h} = \alpha + \beta\, NR_t^K + \varepsilon_{t+h}\,, \qquad K = B \text{ or } G\,,$$

where R_t denotes the current market excess return on the S&P 500 index at time t for "$h = 0$". This setting aligns with a contemporaneous regression framework. For scenarios where $h > 0$, R_{t+h} represents the average excess returns of the market portfolio from $t+1$ to $t+h$ (with h being 1, 3, 6, 9, and 12 months), transitioning the equation into a predictive regression. NR^K represents news ratios. Specifically, NR^B is the monthly proportion of bad news and NR^G represents the proportion of good news. The good or bad news is identified by ChatGPT-3.5. It answers whether the input news is *GOOD* or *BAD* for the stock market. Reported are the regression slopes and R^2s in percentage form. Also reported are the Hodrick (1992) t-statistics. All the forecast variables are standardized to have a zero mean and unit variance. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from January 1996 to December 2022.

	Panel A: Results for *Bad* News			Panel B: Results for *Good* News		
	β (%)	Hodrick-t	R^2 (%)	β (%)	Hodrick-t	R^2 (%)
$h = 0$	-0.58^{*}	-1.82	1.63	0.68^{***}	2.98	2.28
$h = 1$	-0.07	-0.26	0.03	0.44^{*}	1.94	0.96
$h = 3$	0.06	0.21	0.05	0.46^{**}	2.12	3.09
$h = 6$	0.16	0.63	0.65	0.51^{**}	2.01	7.21
$h = 9$	0.25	0.98	2.26	0.50^{*}	1.71	9.82
$h = 12$	0.33	1.25	5.08	0.49	1.61	12.12

Table IA 6. Supplementary Results for Asset Allocation

This table presents the CER gains in percentage points and annualized Sharpe ratio for a mean-variance investor who allocates assets between the market portfolio and risk-free bills, with a risk-aversion coefficient (γ) of one or five. The weight on risky asset is determined by the monthly forecasts of stock market returns based on news ratio (NR^B or NR^G). NR^G (NR^B) is the monthly proportion of good (bad) news identified by ChatGPT-3.5. It answers whether the input news means *GOING UP* or *GOING DOWN* for the stock market. The regression slopes are estimated recursively using the data available at the forecast formation time t. We also consider a proportional transaction cost (TC) of 50bp. The out-of-sample evaluation period is from January 2006 to December 2022.

	CER gain (%)	CER gain (%), TC=50bp	Sharpe Ratio
Panel A: Results for $\gamma = 1$			
NR^B	−1.49	−2.29	0.15
NR^G	5.63	4.73	0.51
Panel B: Results for $\gamma = 5$			
NR^B	−3.21	−3.67	−0.15
NR^G	2.97	1.51	0.53

Table IA 7. Forecasting Macroeconomic Conditions

This table presents the results of following regression,

$$Y_{t+1} = \alpha + \beta\, NR_t^K + \varepsilon_{t+1}\,, \qquad K = B \text{ or } G\,,$$

where Y_{t+1} represents macroeconomic condition variables at future time $t+1$, including the industry production growth (IPG), the VIX of CBOE, the CFNAI, the Aruoba-Diebold-Scotti Business Conditions Index (ADSI), the Kansas City Financial Stress Index (KCFSI), the total non-farm payroll growth (Payroll Growth), the smoothed recession probability, and the real GDP growth (GDPG). NR^K represents news ratios. Specifically, NR^B is monthly proportion of bad news and NR^G is monthly proportion of good news. We identify the good or bad news by using the method of word lists proposed by Loughran and McDonald (2011) or using BERT. Reported are slope estimates (β) and R^2s in percentage form, and also the Newey and West (1987) t-statistics. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from January 1996 to December 2022.

	Panel A: Bad News			Panel B: Good News		
	β	NW-t	R^2 (%)	β	NW-t	R^2 (%)
Word Lists						
IPG	−0.18 **	−2.03	3.20	0.10	1.50	1.07
VIX	0.14	1.58	2.06	−0.10*	−1.66	0.93
CFNAI	−0.17*	−1.83	2.93	0.12*	1.65	1.39
ADSI	−0.17*	−1.73	2.94	0.08*	1.88	0.57
KCFSI	0.20*	1.89	4.07	−0.17***	−2.62	2.80
Payroll Growth	−0.12	−1.19	1.36	0.11	1.31	1.11
SRP	0.27***	2.70	7.10	−0.13**	−2.00	1.82
GDPG	−0.21**	−2.27	4.27	0.17***	2.88	2.83
Bert						
IPG	−0.12	−1.39	1.45	−0.05	−0.86	0.22
VIX	0.12	1.52	1.47	0.17**	2.40	3.02
CFNAI	−0.14	−1.39	1.89	−0.04	−0.65	0.13
ADSI	−0.09	−1.60	0.74	0.04	0.82	0.13
KCFSI	0.12	1.16	1.32	0.05	0.66	0.27
Payroll Growth	−0.10	−0.93	1.05	−0.02	−0.29	0.04
SRP	0.20**	2.17	3.79	0.09	1.34	0.81
GDPG	−0.16**	−2.55	2.40	−0.13	−1.64	1.70